Credit Card Fraud Detection using Hidden Markov Model and Stochastic Tools & technology

Project Report submitted in partial fulfillment of the requirement for the degree of

Master of Technology
in

Computer Science & Engineering

under the Supervision of

Dr. Nitin

By

Shivaca Thakur (132224)

May - 2015

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,
Waknaghat, Solan – 173234, Himachal Pradesh

Certificate

This is to certify that project report entitled "**Credit Card fraud detection using HMM & Stochastic modeling tools & technology** ", submitted by **Shivaca Thakur (132224)** in partial fulfillment for the award of degree of Master of Technology in Computer Science and Engineering to Jaypee University of Information Technology, Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date: **Supervisor's Name**

Dr. Nitin

Acknowledgement

No work is a single man's success. Apart from the hard work and personal efforts, which are key ingredients, it requires the knowledge, guidance, encouragement and support. A work on completion signifies the joint efforts of the persons involved in it.

Firstly, I take this opportunity to express a deep sense of gratitude towards my guide **Dr. Nitin** for providing excellent guidance, encouragement and inspiration throughout the project work. Without his invaluable guidance, this work would never have been a successful one.

I wish to express my gratitude and high regards to **Prof. Dr. Satya Prakash Ghrera** head of CSE department, JUIT.

I would like to express my profound sense of gratitude to all the faculty members for their cooperation and encouragement throughout my course.

Last but not the least; I would also like to thank my family and friends for their valuable suggestions and undue support which has been a constant source of encouragement in all my educational pursuits.

Signature of the student

Name of Student: Shivaca Thakur (132224)

Date

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AVS	Address Verification Systems
CVV	Card Verification Value Code
ATM	Automatic Teller Machine
HMM	Hidden Markov Model
FPMSC	Fraud Patterns Mining Service Center
XML	Extensible Markup Language
SOAP	Simple Object Access Protocol
HTTP	Hypertext Transfer Protocol
SMTP	Simple Mail Transfer Protocol
FPM	Fraud Patterns Mining
PMML	Predictive Model Markup Language
FDS	Fraud Detection System
NB	Naïve Bayesian
BP	Back Propagation
SBT	Suspicion Building Tool
SOM	Self-Organizing Map
ANN	Artificial Neural Network
ASPECT	Advanced Security for Personal Communications Technologies
GUI	Graphical User Interface
GCM	Global Constants Module
GUIM	Core/Graphical User Interface Module
DBIM	Database Interface Module
LAL	Learning Algorithms Library
DLL	Dynamic Link Library
GA	Genetic Algorithm
CC	Compute Critical values
SQL	Structured Query Language
SVM	Support Vector Machine

Abstract

The advent of credit card increases the people comfort but also attracts fraudsters. Credit cards are good targets for fraud, because in a short time large amount of money can be earned without taking risks. The crime will be discovered after few weeks so it is easy for malicious agents to commit this crime. For the past 20 years financial organizations have seen increase in the amount and types of fraud.

The best method is to testify the reasons of fraud from the available data. From several researches the solutions for this credit card fraud are determined by genetic algorithms, artificial intelligence, artificial immune systems, visualization, database, behavioral, distributed and parallel computing, fuzzy logic, neural networks and pattern recognition. There are many specialized fraud detection solutions which protects credit card, insurance, retail, telecommunications industries. The main objective of these detection systems is to identify the trends of fraudulent transactions.

Out of these techniques, we chose HMM and Stochastic (behavioral), and compared the two in terms of the detection of frauds. Stochastic proves to be better in terms of accuracy.

Chapter 1

Introduction

Chapter 1

Introduction

These days internet has become an important part of our life. A person can do shopping, investments, and perform all the banking tasks online whenever he wants. Almost, all the organizations have their own website, where customer can perform all the tasks like shopping. They just have to provide their credit card details. Online banking and e-commerce organizations have been experiencing an increase in the credit card transactions and other modes of on-line transactions. [1]

Fraud can be defined as wrongful or criminal deception which aims to financial or personal gain. The two main mechanisms to avoid frauds and losses due to fraudulent activities are fraud prevention and fraud detection systems. Fraud prevention is the mechanism with the goal of preventing the occurrence of fraud. Fraud detection system comes into play when the fraudsters surpass the fraud prevention systems and start a fraudulent transaction. [2]

In the present electronic society, e-commerce has become an essential sales channel for global business and development. Due to rapid advancement in e-commerce, use of credit cards for purchases has been increased dramatically. Unfortunately, fraudulent use of credit cards has also become an attractive source of revenue for criminals. Occurrence of credit card fraud is increasing by a high rate due to the exposure of security weaknesses in traditional credit card processing systems which results in the loss of a huge amount of money every year. Fraudsters now use refined techniques to commit credit card fraud.

Worldwide, the fraudulent activities present some unique challenges to banks and other financial institutions who issue credit cards to the customers. In case of bank cards (Visa and MasterCard) a study done by American Bankers Association in 1996 reveals that the estimated gross fraud loss was $790 million in 1995. [3] The majority of the loss due to credit card fraud is suffered by the USA alone.

This is not surprising since 71% of all credit cards in the world are issued in USA only. In 2005, the total fraud loss in the USA was reported to be $2.7 billion and it has increased up to $3.2 billion in 2007. [4] Another survey done on over 160 companies revealed that online fraud is committed 12 times higher than offline fraud. [5]

To address this problem, financial institutions employ various fraud prevention tools like real-time credit card authorization, address verification systems (AVS), card verification codes, rule based detection, etc. [6]. But fraudsters are adaptive, and given time, they devise several ways to circumvent such protection mechanisms.

Despite the best efforts of the financial institutions, law enforcement agencies and the government, still credit card fraud continues to increase. In addition to the significant financial losses, the main concern of the law enforcement agencies is that worldwide this money can also be used to support other criminal activities.

Thus, once the fraud prevention measures have failed, there is a need for to detect fraud in order to maintain the capability of the payment system using some effective technologies. Fraudsters constitute a very inventive and fast moving fraternity. With time as the preventive technology changes, so does the technology used by the criminals and the ways they go about with their fraudulent activities.

Purchases using by the credit cards are two types. They are classified as: [7]

1. **Physical card:**

   In this physical card purchased system, the cardholder gives his card physically to the vendor to make a payment. To carry out false transactions in this type of purchase, the attacker has to steal the credit card.

   The credit card companies have to face a huge loss of money if the cardholder does not know that he has lost his card.

2. **Virtual card:**

In the second type of the card that i.e. Virtual card purchase system, we need to know some important information about the card holder like the card number, security code, the issue date and the expiry date. All these details are required to make the payment. We can use these types of purchases when purchasing is done online or payments are made by the phone like mobile top-ups etc.

To commit fraud in these types of purchases, the fraudster has to know all the details of the card holder. It is not an easy task to do so. Most of the time, the real cardholder does not that someone else has stolen his card details and is using them. The only way to detect this kind of fraud is to observe the spending patterns on every card and to notice any kind of irregularity with respect to the "usual" spending patterns.

Fraud detection is mainly based on the analysis of the purchases which the card holder have completed like some online shopping or payment of bills. The cardholder keeps the receipt of the purchase and for purchasing something next time. Then by comparing both the values, the card holder can easily recognize what is the last amount and the present amount.

Since humans tend to demonstrate explicit behaviorist profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc. Deviation from such patterns represents a potential threat to the system.

There are many previous studies done on credit card fraud detection. The most commonly used fraud detection methods in this domain are rule-induction techniques, decision trees, Artificial Neural Networks (ANN), Support Vector Machines (SVM), logistic regression, and meta-heuristics such as genetic algorithms.

These techniques can be used alone or in collaboration using ensemble or meta-learning techniques to build classifiers. Most of the credit card fraud detection systems use supervised algorithms such as neural networks; decision tree techniques like ID3, C4.5 and C&RT; and SVM. [2]

1.1. Credit Card Fraud

Credit card fraud is a wide-ranging term for theft and fraud committed involving a payment car (credit card or debit card), as a fraudulent source of resources in a transaction. The purpose is to get goods without making a payment, or to obtain unauthorized money from an account. Credit card fraud is also an addition to identity theft. According to the United States Federal Trade Commission, while identity theft had been holding stable for the last few years, it saw a 21 percent increase in 2008. However credit card fraud, the crime which most people associate with ID theft, decreased as a percentage of all ID theft complaints. [8]

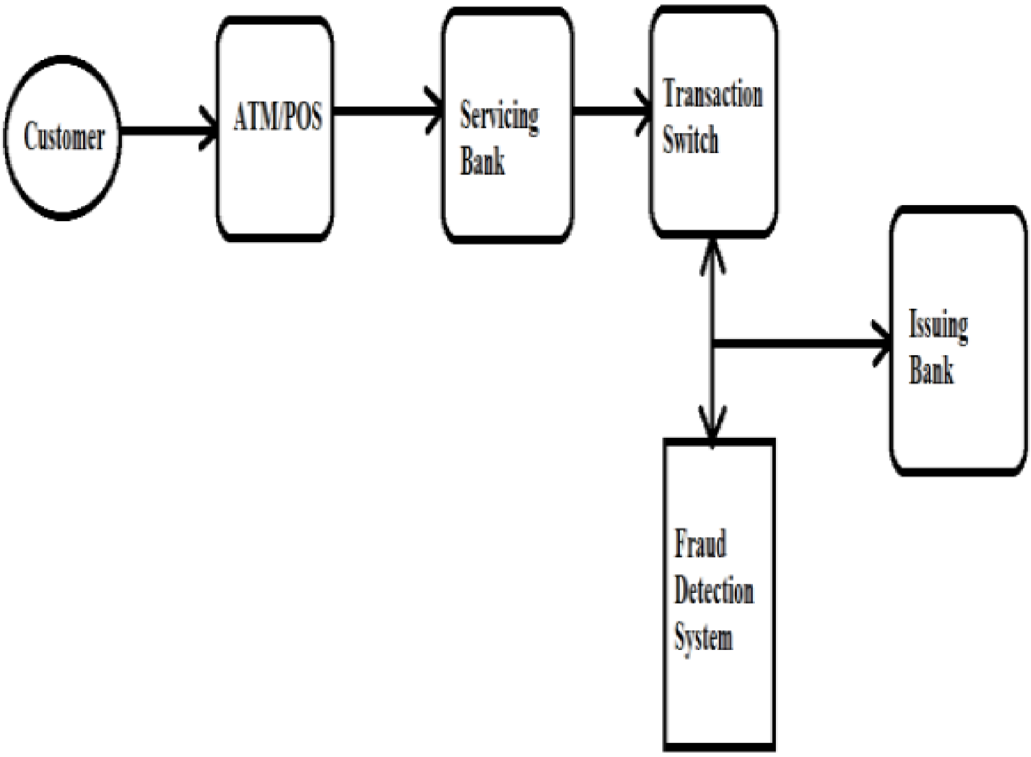


Figure 1.1 Basic credit card fraud detection system

Card fraud begins either with the theft of the physical card or with the compromise of data associated with the account, including the card account number or other information that would routinely and necessarily be available to a card holder during a genuine transaction. The compromise can occur by many common routes and can usually be conducted without tipping off the card holder, the issuer, at least until the account is ultimately used for fraud.

A simple example is that of a store clerk copying sales receipts for using it later on. The fast growth of credit card use on the Internet has made database security lapses particularly costly; in some cases, millions [9] of accounts have been compromised.

Stolen cards can be reported quickly by cardholders, but a compromised account can be accumulated by a thief for weeks or months before any fraudulent use which makes it difficult to identify the source of the compromise. The cardholder may not discover fraudulent use until receiving a billing statement, which may be delivered infrequently. Cardholders can decrease this fraud risk by checking their account frequently to ensure constant awareness in case there are any suspicious, unknown activities or transactions.

1.2. Types of Frauds

Credit card security relies on the physical security of the plastic card as well as the privacy of the credit card number. CVV (Card Verification Value) code is a new authentication procedure introduced by credit card companies to reduce fraud in the internet transactions.

Credit card fraudsters employ a large number of techniques to commit fraud. In credit card business, fraud occurs when a lender is fooled by a borrower by offering purchases, believing that the borrower credit card account will provide payment for this purchase. Ideally, no payment will be made. If the payment is made, the credit card issuer will reclaim the amount paid. Fraudsters can either internal party or external party.

As an external party, fraud is committed being a prospective/existing customer or a prospective/existing supplier. To combat the credit card fraud effectively, it is important to first understand the different types of credit card fraud**.**

1.2.1. Card not present transaction

The mail and the Internet are major routes for fraud against merchants who sell and ship products, and affects legitimate mail-order and Internet merchants. If the card is not physically present (called CNP, card not present) the merchant must rely on the holder (or someone purporting to be so) presenting the information indirectly, whether by mail, telephone or over the Internet. While there are safeguards to this, [10] it is still more risky than presenting in person, and indeed card issuers tend to charge a greater transaction rate for CNP, because of the greater risk.

It is difficult for a merchant to verify that the actual cardholder is indeed authorising the purchase. Shipping companies can guarantee delivery to a location, but they are not required to check identification and they are usually not involved in processing payments for the merchandise. A common recent preventive measure for merchants is to allow shipment only to an address approved by the cardholder, and merchant banking systems offer simple method of verifying this information.

Before this and similar counter measures were introduced, mail order carding was rampant as early as 1992. [11] A carder would obtain the credit card information for a local resident and then intercept delivery of the illegitimately purchased merchandise at the shipping address, often by staking out the porch of the residence.

Small transactions generally undergo less scrutiny, and are less likely to be investigated by either the card issuer or the merchant. CNP merchants must take extra precaution against fraud exposure and associated losses, and they pay higher rates for the privilege of accepting cards. Fraudsters bet on the fact that many fraud prevention features are not used for small transactions.

Merchant associations have developed some prevention measures, such as single use card numbers, but these have not met with much success. Customers expect to be able to use their credit card without any hassles, and have little incentive to pursue additional security due to laws limiting customer liability in the event of fraud. Merchants can implement these prevention measures but risk losing business if the customer chooses not to use them.

1.2.2. Identity Theft

Identity theft can be divided into two broad categories: application fraud and account takeover.

1.2.2.1. Application Fraud

Application fraud takes place when a person uses stolen or fake documents to open an account in another person's name. Criminals may steal documents such as utility bills and bank statements to build up useful personal information. Alternatively, they may create fake documents. With this information, they could open a credit card account or loan account in the victim's name, and then fully draw it.

1.2.2.2. Application Takeover

Account takeover takes place when a person takes over another person's account, first by gathering personal information about the intended victim, and then contacting their card issuer while impersonating the genuine cardholder, and asking for mail to be redirected to a new address. The criminal then reports the card lost and asks for a replacement card to be sent. They may then set up a new PIN. They are then free to use the card until the rightful cardholder discovers the deception when he or she tries to use their own card, by which time the account would be drained.

1.2.3. Skimming

Skimming is the theft of payment card information used in an otherwise legitimate transaction. The thief can procure a victim's card number using basic methods such as photocopying receipts or more advanced methods such as using a small electronic device (skimmer) to swipe and store hundreds of victims' card numbers. Common scenarios for skimming are restaurants or bars where the skimmer has possession of the victim's payment card out of their immediate view. [12]

The thief may also use a small keypad to unobtrusively transcribe the 3 or 4 digit Card Security Code, which is not present on the magnetic strip. Call centres are another area where skimming can easily occur. [13] Skimming can also occur at merchants such as gas stations when a third-party card-reading device is installed either outside or inside a fuel dispenser or other card-swiping terminal. This device allows a thief to capture a customer's card information, including their PIN, with each card swipe. [14]

Skimming is difficult for the typical cardholder to detect, but given a large enough sample, it is fairly easy for the card issuer to detect. The issuer collects a list of all the cardholders who have complained about fraudulent transactions, and then uses data mining to discover relationships among them and the merchants they use. For example, if many of the cardholders use a particular merchant, that merchant can be directly investigated. Sophisticated algorithms can also search for patterns of fraud.

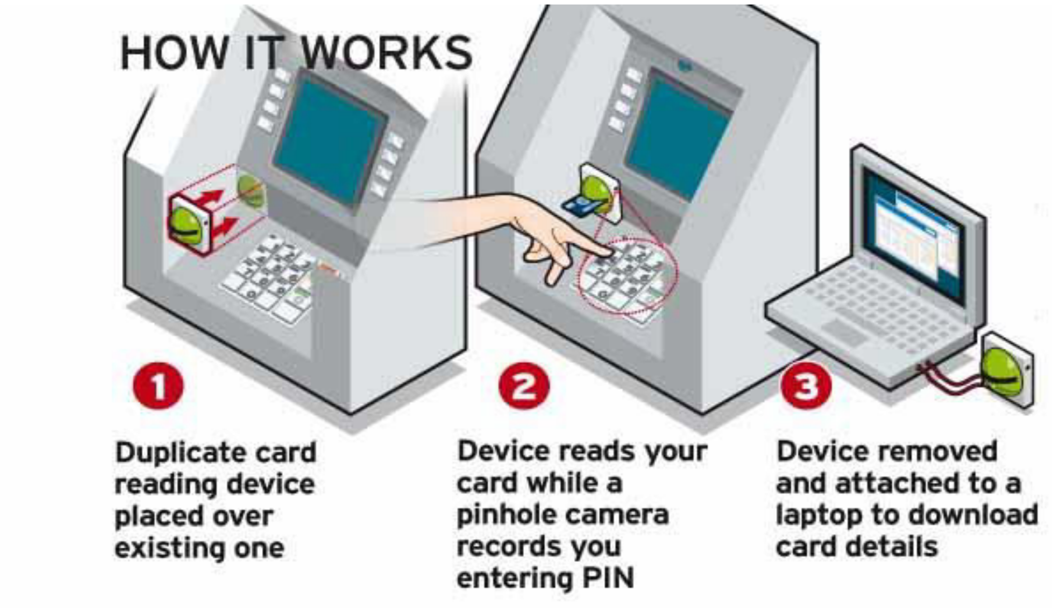

HOW IT WORKS
1
Duplicate card reading device placed over existing one
2
Device reads your card while a pinhole camera records you entering PIN
3
Device removed and attached to a laptop to download card details

Figure 1.2 How skimming works

Merchants must ensure the physical security of their terminals, and penalties for merchants can be severe if they are compromised, ranging from large fines by the issuer to complete exclusion from the system, which can be a death blow to businesses such as restaurants where credit card transactions are the norm.

1.2.4. Carding

Carding is a term used for a process to verify the validity of stolen card data. The thief presents the card information on a website that has real-time transaction processing. If the card is processed successfully, the thief knows that the card is still good. The specific item purchased is immaterial, and the thief does not need to purchase an actual product; a web site subscription or charitable donation would be sufficient. The purchase is usually for a small monetary amount, both to avoid using the card's credit limit, and also to avoid attracting the card issuer's attention. A website known to be susceptible to carding is known as a cardable website.

In the past, carders used computer programs called "generators" to produce a sequence of credit card numbers, and then test them to see which the valid accounts were. Another variation would be to take false card numbers to a location that does not immediately process card numbers, such as a trade show or special event.

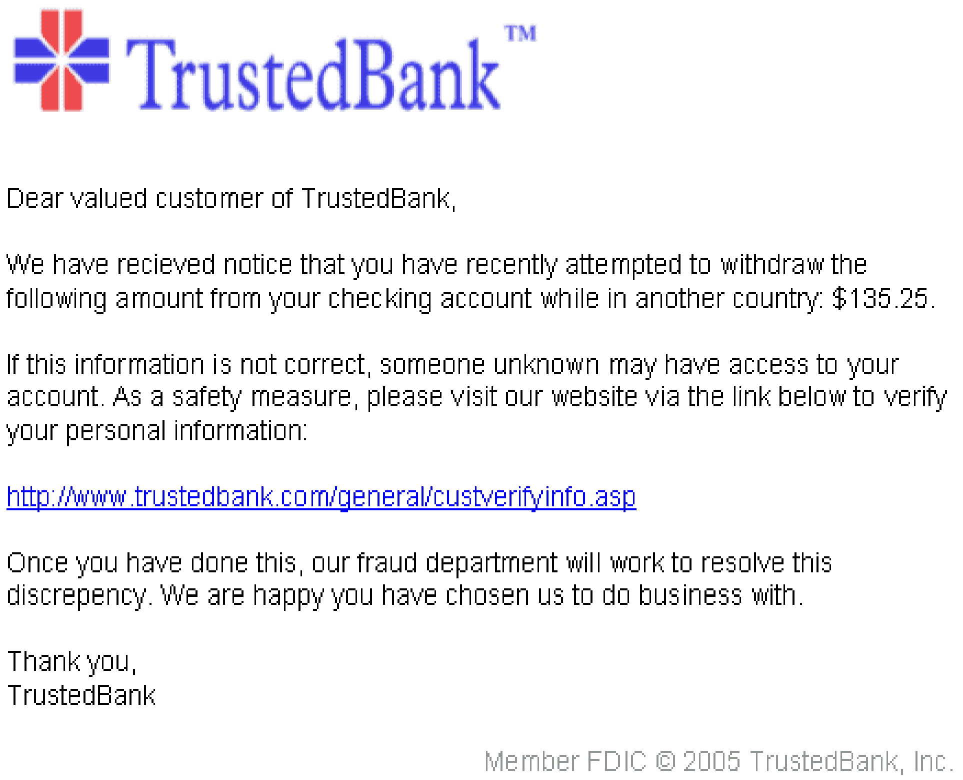
However, this process is no longer viable due to widespread requirement by internet credit card processing systems for additional data such as the billing address, the 3 to 4 digit Card Security Code and/or the card's expiration date, as well as the more prevalent use of wireless card scanners that can process transactions right away. Nowadays, carding is more typically used to verify credit card data obtained directly from the victims by skimming or phishing.

1.2.5. BIN Attack

Credit cards are produced in BIN ranges. Where an issuer does not use random generation of the card number, it is possible for an attacker to obtain one good card number and generate valid card numbers by changing the last four numbers using a generator. The expiry date of these card IDs would most likely be the same as the good card.

1.2.6. Phishing

An example of a phishing email, disguised as an official email from a (fictional) bank. The sender is attempting to trick the recipient into revealing confidential information by "confirming" it at the phisher's website. Note the misspelling of the words received and discrepancy. Also note that although the URL of the bank's webpage appears to be legitimate, the hyperlink would actually be pointed at the phisher's webpage.



TrustedBank™

Dear valued customer of TrustedBank,

We have recieved notice that you have recently attempted to withdraw the following amount from your checking account while in another country: $135.25.

If this information is not correct, someone unknown may have access to your account. As a safety measure, please visit our website via the link below to verify your personal information:

http://www.trustedbank.com/general/custverifyinfo.asp

Once you have done this, our fraud department will work to resolve this discrepency. We are happy you have chosen us to do business with.

Thank you,
TrustedBank

Member FDIC © 2005 TrustedBank, Inc.

Fig 1.3 An Example of phishing

Above is an example of a phishing email, disguised as an official email from a (fictional) bank. The sender is attempting to trick the recipient into revealing confidential information by "confirming" it at the phisher's website. Note the misspelling of the words received and discrepancy. Also note that although the URL of the bank's webpage appears to be legitimate, the hyperlink would actually be pointed at the phisher's webpage.

What a typical "phishing" e-mail may look like. The bank here is fictional, but it is to be assumed that a real phishing attempt would claim to be from an actual bank the customer belongs to. Notice how it tries to establish authenticity by using the bank's logo and providing what appears to be link to a website the customer has been to many times before. This mock-up was created by me on December 2, 2005 and placed in the public domain. Note the effect achieved by not using "i before e, except after c", thus misspelling the word "received".

1.2.6.1.Tele Phishing

Scammers may obtain a list of individuals with their name and phone number luring victims into thinking that they are speaking with a trusted organization handing over sensitive information such as credit card details. Scamming has moved from landlines to cellphones in recent years. One popular tactic is to claim that they are from the “Card Services” division of one, or any number of popular banks, and are “verifying” your account information so that they can provide you a lower interest rate. Scammers can be very convincing, aggressive, and tireless in their efforts, often organized into large but clearly mobile call centers.

1.2.7. Balance Transfer Checks

Some promotional offers include active balance transfer checks which may be tied directly to a credit card account. These are often sent unsolicited, and may occur as often as once per month by some financial institutions. In cases where checks are stolen from a victim’s mailbox they can be used at point of sales locations thereby leaving the victim responsible for the losses. They are one path at times used by fraudsters.

1.3. Detecting and Preventing Frauds [15]

1.3.1. Preventing Bankruptcy Fraud using Credit Bureaus

Bankruptcy fraud is one of the most difficult types of fraud to predict. Bankruptcy fraud means that purchasers use credit cards knowing that they are not able to pay for their purchases. The bank will send them an order to pay.

However, the customers will be recognized that they are not able to recover their debts. The only way to prevent this bankruptcy fraud is by doing a pre-check with credit bureau. Information in the credit bureau data is gathered from many different sources.

Banks, consumer finance companies, credit unions, and collection agencies are some of the entities that periodically report to the credit bureau. Data are also obtained from state and federal courts on judgments, liens, and bankruptcy filings.

The process is as follows:

- The bank passes an enquiry to the credit bureau, who uses a third party to gather information. The enquiry includes identification information required by the credit bureau.
- The credit bureau sends a credit report for this single individual including personal particulars, details of non-compliance with contractual obligations, information from public directories and additional positive information such as repayment of loans according to contract at or before maturity. Some credit bureaus are also able to trace the address of a specific individual, who has moved to an ‘unknown’ address.
- A credit file is created when an individual applies for, or uses, credit or a public record & is reported to the credit bureau.
- Once a credit file is established, consumer’s credit-seeking behavior, payment and purchase behavior, and any changes to the public records are recorded to estimate, detect, or avoid undesirable behavior and the updates are posted.
- Once the bank has received the credit report from the credit bureau, the bank can identify insolvency cases.

1.3.2. Detecting Charge-backs through over limit/Vintage Reports

Theft fraud means using a card that is not owned by him/her. The fraudster will steal the card of someone else and use it as many times as much as possible before the card is blocked. The owner must react and contact the bank sooner, so that, bank will take measures to stop the fraudster.

Counterfeit fraud occurs when the credit card is used remotely and only the credit card details are needed. The fraudster will copy your card number and codes and use it via certain web-sites, where no signature or physical cards are required. Fraudsters use credit card data which is stolen and the merchant faces money loss and this is named as "charge-backs". Charge-backs are generated if credit card holders object to items on their monthly credit card statements.

This type of fraud can be detected through ‘over limit’ reports or ‘vintage’ reports. These reports provide a daily list of customers that have exceeded their credit limit. A certain degree of tolerance may be accepted. For the credit card listed, the customers are contacted and if they do not react, the card is blocked. ATM transactions of large amounts and purchases of goods for a larger amount than normal are suspicious and must be notified to the customer.

1.3.3. Detecting Duplicates/Identity Fraudsters using Cross-matching technique

Someone applies for a credit card with false information is said to be Application Fraud.

Two modes of application fraud are:

- Duplicates, and
- Identity fraudsters.

When applications come from a same individual with same details, it is called as duplicates. When applications come from different individuals with similar details is called as identity fraudsters.

The bank requires some details from the credit card applicants such as identification information, location information, contact information, confidential information and additional information. All these characteristics may be used individuals with more than one card can be identified.

In contrast, identity fraudster is perpetrated by real criminals for searching duplicates. Cross-matching technique is used to identify the duplicates and identity fraudsters. To detect the duplicates simple queries that give fast results are passed to cross-identify the information with location details.

Identity fraudster may be either identity fraud (contain plausible) or identity theft (real but stolen identity information). Many matching rules must be applied and it is acknowledged that many false positive cases will be identified.

Chapter 2

Literature Review

Chapter 2
Literature Review

2.1. Credit Card Fraud Detection using Hidden Markov Model [16]

2.1.1. HMM (Hidden Markov Model)

An HMM is a double embedded stochastic process with two hierarchy levels. It can be used to model much more complicated stochastic processes as compared to a traditional
Markov model.

An HMM has a finite set of states governed by a set of transition probabilities. In a particular state, an outcome or observation can be generated according to an associated probability distribution. It is only the outcome and not the state that is visible to an external observer. [17]

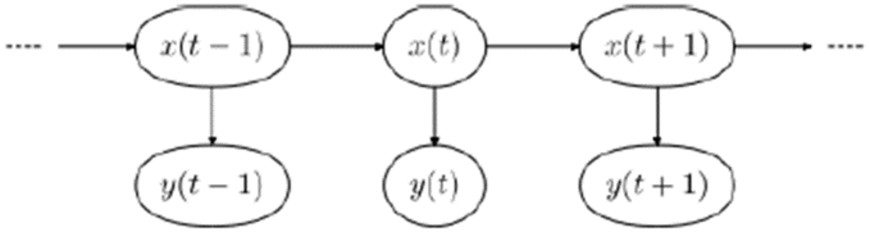


Fig 2.1 Architecture of HMM

HMM-based applications are common in various areas such as speech recognition, bioinformatics, and genomics. In recent years, Joshi and Phoba [18] have investigated the capabilities of HMM in anomaly detection. They classify TCP network traffic as an attack or normal using HMM. Cho and Park [19] suggest an HMM-based intrusion detection system that improves the modeling time and performance by considering only the privilege transition flows based on the domain knowledge of attacks. Ourston et al. [20] have proposed the application of HMM in detecting multistage network attacks. Hoang et al. [21] present a new method to process sequences of system calls for anomaly detection using HMM.

The key idea is to build a multilayer model of program behaviors based on both HMMs and enumerating methods for anomaly detection. Lane [22] has used HMM to model human behavior. Once human behavior is correctly modeled, any detected deviation is a cause for concern since an attacker is not expected to have a behavior similar to the genuine user. Hence, an alarm is raised in case of any deviation. An HMM can be characterized by the following [22]: The diagrams (figure 2.1 & 2.2) below shows the general architecture of an instantiated HMM.

Each oval shape represents a random variable that can adopt any of a number of values. The random variable x(t) is the hidden state at time t (with the model from the above diagram, x(t) ∈ { x1, x2, x3 }). The random variable y(t) is the observation at time t (with y(t) ∈ { y1, y2, y3, y4 }). The arrows in the diagram (often called a trellis diagram) denote conditional dependencies.

From the Fig. 2.1, it is clear that the conditional probability distribution of the hidden variable x(t) at time t, given the values of the hidden variable x at all times, depends only on the value of the hidden variable x(t − 1), the values at time t − 2 and before have no influence. This is called the Markov property. Similarly, the value of the observed variable y(t) only depends on the value of the hidden variable x(t) (both at time t).

In the standard type of hidden Markov model considered here, the state space of the hidden variables is discrete, while the observations themselves can either be discrete (typically generated from a categorical distribution) or continuous (typically from a Gaussian distribution). The parameters of a hidden Markov model are of two types, transition probabilities and emission probabilities (also known as output probabilities). The transition probabilities control the way the hidden state at time t is chosen given the hidden state at time t − 1.

The hidden state space is assumed to consist of one of N possible values, modeled as a categorical distribution. This means that for each of the N possible states that a hidden variable at time t can be in, there is a transition probability from this state to each of the N possible states of the hidden variable at time t + 1, for a total of N2 transition probabilities.

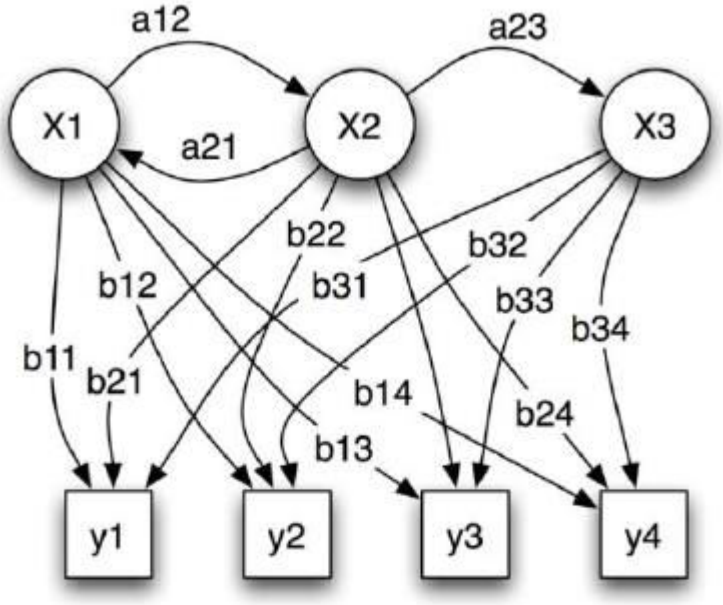


Fig 2.2 Another architecture of HMM

2.1.2. Algorithm

Step 1: Generate the synthetic data according to given Probability. Use to separate distribution for Genuine and Fraud transactions.

Step 2: Read the generated data.

Step 3: Re-categorize the data into five groups as transaction month, date, day, amount of transaction & difference between successive transaction amounts.

Step 4: Make each transaction data as vector of five fields.

Step 5: Make two separate groups of data named True & False transaction group (if false transaction data is not available add randomly generate data in this group).

Step 6: Train HMM.

Step 7: Save the trained matrix.

Step 8: Read the current Transaction.

Step 9: Repeat the process from step3 for current transaction data only.

Step 10: Place the saved Matrix & currently generated vector in classifier.

Step 11: Take the generated decision from the classifier.

2.1.3. Advantages

- Provide decision support system to prevent frauds and control risks
- Overcomes the problem of high false alarm rate
- Accuracy up to 83%
- Negligible delay

2.1.4. Disadvantages

- Not suitable for outlier checking and comprehensive evaluating

2.2. A Web Services-Based Collaborative Scheme for Credit Card Fraud Detection [23]

2.2.1. Architecture

In the web services-based collaborative detection scheme, participant banks plays as service consumers, while Fraud Patterns Mining Service Center (FPMSC) serves as the service provider.

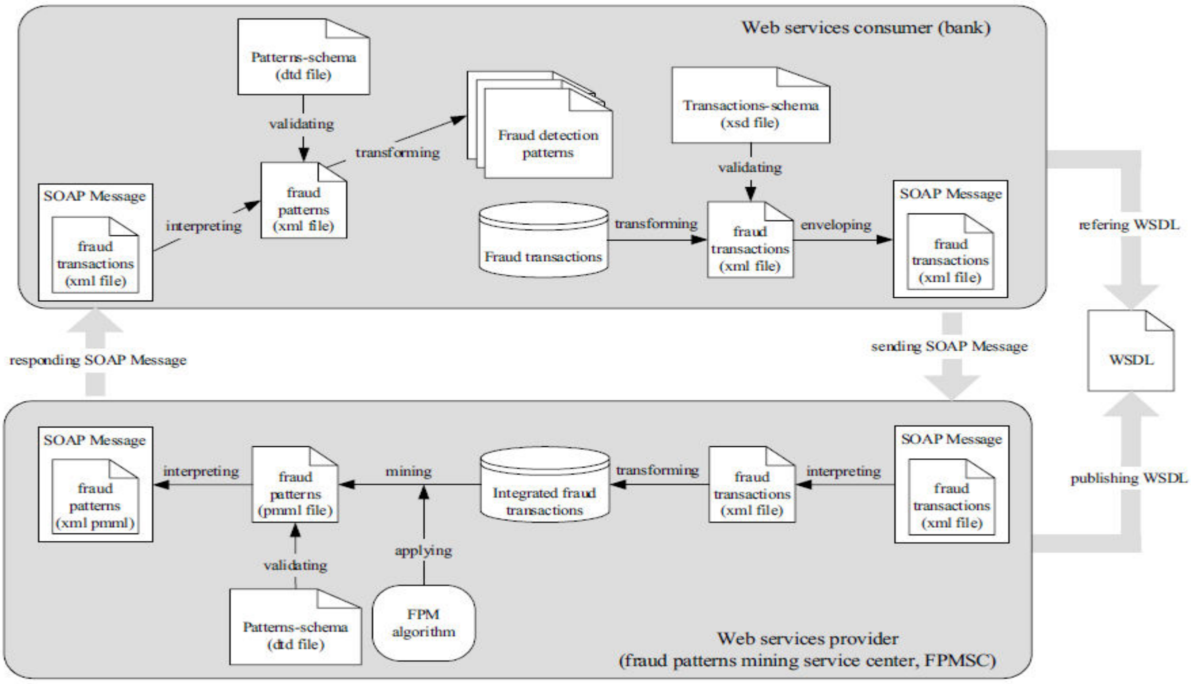

Web services consumer (bank)
Patterns-schema (dtd file)
validating
transforming
Fraud detection patterns
Transactions-schema (xsd file)
validating
SOAP Message
fraud transactions (xml file)
interpreting
fraud patterns (xml file)
Fraud transactions
transforming
fraud transactions (xml file)
enveloping
SOAP Message
fraud transactions (xml file)
refering WSDL
responding SOAP Message
sending SOAP Message
WSDL
publishing WSDL
SOAP Message
fraud patterns (xml pmml)
interpreting
fraud patterns (pmml file)
mining
Integrated fraud transactions
transforming
fraud transactions (xml file)
interpreting
SOAP Message
fraud transactions (xml file)
validating
applying
Patterns-schema (dtd file)
FPM algorithm
Web services provider
(fraud patterns mining service center, FPMSC)

Fig 2.3 Architecture of Web Services Based Collaborative Scheme for Fraud Detection

To achieve data exchange across heterogeneous applications of banks, participant banks must obey uniform data formats for validating the exchanged data. FPMSC publishes a WSDL file that describes the implementation and interface specification of its provided service. Banks must obey the regulations in the WSDL file, so that they can know what data should be sent and what patterns will be replied, and understand how to access the service. In the WSDL file, the input message of provided service is defined as fraud transactions sent from banks, while the output message is defined as fraud patterns replied to banks.

Furthermore, FPMSC defines the specific schemas for input and output messages to ensure the contents of input and output messages are valid. The specific schema for input message, called Transactions-schema, is specified based on XML Schema [24]. The specific schema for output message, called Patterns-schema, is specified based on PMML standard [25]. PMML is a markup language based on DTD standard [26] for defining the predictive models produced by data mining systems.

When participant banks want to access the collaborative fraud patterns mining service provided by FPMSC, they must transform their individual fraud transactions stored in legacy formats to an XML document that must pass through the validation by

Patterns-schema. The valid xml document is enveloped in SOAP Envelope within SOAP Message. The SOAP Message can be sent to FPMSC via popular protocols such as HTTP, SMTP, and MIME.

The fraud transactions enveloped in SOAP Message sent to FPMSC can be accumulated into the integrated fraud transactions. FPMSC extracts fraud patterns from the integrated fraud transactions using Fraud Patterns Mining (FPM) algorithm. FPM algorithm is developed based on Apriori algorithm [27] for mining fraud pattern association rules which manifest the information about what features exist in popular fraud transactions. The details of FPM algorithm are introduced in next section. The uncovered fraud patterns are transformed to a PMML document validated by Patterns-schema. The valid PMML document is then enveloped in SOAP Envelope within SOAP Message. Via the same protocols, the SOAP Message is replied to the bank that accesses the service.

When receiving the SOAP Message sent from FPMSC, participant banks can interpret the contents of PMML document within SOAP Message based on Patterns-schema and retrieve fraud patterns. Through those fraud patterns, banks can enhance their original fraud detection systems to avoid suffering fraud attacks.

2.2.2. Fraud patterns mining algorithm

2.2.2.1. Discretization for continuous attributes

Let T be the integrated fraud transactions provided by banks. A fraud transaction t, t $\in$T, contains n attributes. Some attributes are discrete, and others are continuous. The purpose of discretization is to divide values of a continuous attribute into several discrete intervals, so that each interval can be regarded as a discrete value of the attribute. FPM algorithm quantizes continuous attributes based on the merging and unsupervised concepts.

Let A be a continuous attribute, and a_g is a continuous value of A. Initially, all continuous values of A are divided into k equal width intervals where k is the number of intervals.

All adjacent intervals are evaluated to find the best pair of adjacent intervals to be merged. The number of intervals should reduce one after the merge operation. The discretization process continues until the stopping criterion is satisfied.

$$D(I_d) = \frac{n_d}{u_d - l_d}$$

(1)

$$Avg(I_d) = \frac{\sum a_g}{n_d} \text{ where } l_d \le a_g < u_d$$

(2)

$$\Delta D(I_d, I_{d+1}) = | D(I_d) - D(I_{d+1}) |$$

(3)

$$\Delta Avg(I_d, I_{d+1}) = | Avg(I_d) - Avg(I_{d+1}) |$$

(4)

$$Dissim(I_d, I_{d+1}) = \Delta D(I_d, I_{d+1}) \times \Delta Avg(I_d, I_{d+1})$$

(5)

I_d and I_{d+1} are two adjacent intervals.

$D(I_d)$ is the density of I_d

n_d is the number of continuous values belonging to A

l_d/u_d are the lower/upper bound

$Avg(I_d)$ is the average of continuous values belonging to I_d

a_g is a continuous value of A

The discretization process recursively executes with Equations (1), (2), (3), (4), and (5) until the stopping criterion is satisfied. The stopping criterion is satisfied if either Equation (6) or Equation (7) is true.

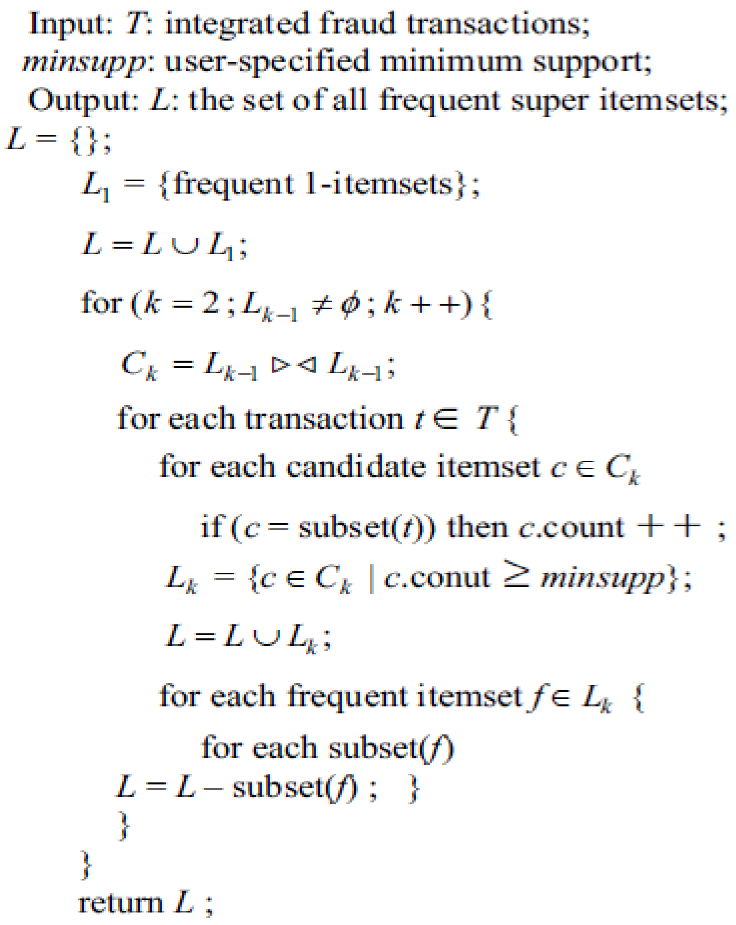
$$\Delta D(I_d, I_{d+1}) > \frac{1}{2} \times [D(I_d) + D(I_{d+1})] \ \forall I_d, I_{d+1} \in A$$

(6)

$$\Delta Avg(I_d, I_{d+1}) > \frac{1}{2} \times (l_d + u_{d+1}) \ \forall I_d, I_{d+1} \in A$$

(7)

2.2.2.2. Mining of Fraud Patterns



```
 Input: $T$: integrated fraud transactions;
 $minsupp$: user-specified minimum support;
 Output: $L$: the set of all frequent super itemsets;
$L = \{\}$;
    $L_1 = \{$frequent 1-itemsets$\}$;
    $L = L \cup L_1$;
    for ($k = 2$ ; $L_{k-1} \neq \phi$ ; $k$ + +) {
      $C_k = L_{k-1} \bowtie L_{k-1}$;
      for each transaction $t \in T$ {
        for each candidate itemset $c \in C_k$
          if ($c$ = subset($t$)) then $c$.count + + ;
        $L_k = \{c \in C_k \mid c.\text{conut} \geq minsupp\}$;
        $L = L \cup L_k$;
        for each frequent itemset $f \in L_k$ {
          for each subset($f$)
      $L = L -$ subset($f$) ;   }
      }
    }
    return $L$ ;
```

2.2.3. Advantages

- The original fraud detection system of the bank is developed based on decision tree techniques. The system can enhance its detection ability by adjusting the weights of induction rules according to the uncovered fraud patterns.

- Since the fraud patterns are represented in the form of rules, it is impossible for other banks to decode original fraud transactions sent from the bank.

2.2.4. Disadvantages

- This technique cannot used to detect new frauds.

2.3. Game-Theoretic Approach to Credit Card Fraud Detection [20]

Consider a database system in an organization and a set of authorized users who have access rights on the database such as in banking services, credit card companies, etc. There always exists the possibility of legitimate and even non-legitimate transactions, what we hereby term as fraudulent transactions, being attempted by the authorized users or more typically, by adversaries posing as authorized users.

The primary objective of any defense mechanism monitoring such an application would be to identify these fraudulent transactions as early as possible while limiting the possibility of raising too many false alarms. This form of Intrusion Detection in databases is an essential component of Information Warfare.

The situation can be visualized as two adversaries playing against each other, the attacker launching attacks against the database system and the detection system countering it. The problem effectively models as a typical game with each player trying to outdo the other and Game theory has long been used to tackle such problems.

2.3.1. Fraud Detection System

2.3.1.1.Game-Theoretic Model

The presence of two parties with conflicting goals provided us with the initial impetus to use Game theory as an approach for fraud detection.

The game, in case of three transaction ranges, can be modeled as shown in Figure 2.4. The thief, oblivious of the ranges or the strategies used by the FDS, needs to choose the i^{th} range from the possible ’n’ ranges. The FDS, in contrast, is unaware of the thief’s choice and hence, the possible choices form the information set for the FDS. A correct prediction of the i^{th} range by the FDS results in the thief being caught.

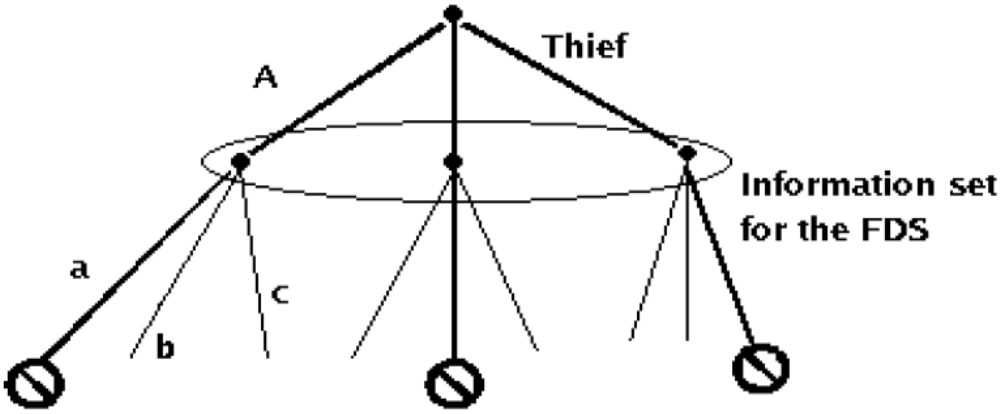


Fig 2.4. Modeling the Game

2.3.1.2. Architecture

The Fraud Detection System comprises of two layers, the ’Rule-based component’ and the ’Game-theoretic component’.

The First Layer:

The first layer should not only include certain features from available systems but also because we do not want to tackle millions of transactions with the Game theory rules, most of which are carried out due to routine use of credit cards.

This layer would have rules like average daily/ monthly buying, shipping address being different from billing address, etc. In addition, customer-specific rules can also be incorporated. Intuitively, the first layer can filter out seemingly genuine transactions as is being done by the existing systems.

The First Layer flags a transaction as ’suspect’ if it crosses a user-defined threshold level. This introduces a trade-off between false positives (when the threshold is low) and more seriously, false negatives (when the threshold is high). We introduce the Second Layer in order to tackle this issue.

The Second Layer:

The second tier is the Game-theoretic component of the model. We consider the game between the fraudster and the FDS to be a multistage repeated game. This is essential because, firstly, the fraudster is likely to try again even if he fails with one card and secondly, no effective learning can take place if the game is considered to be a one-shot one.

The game being played between the FDS and the fraudster is one of incomplete information since the fraudster would be completely unaware of the modus operandi of the Detection System. However, the fraudster is likely to have some notions or beliefs about the strategy of the FDS.

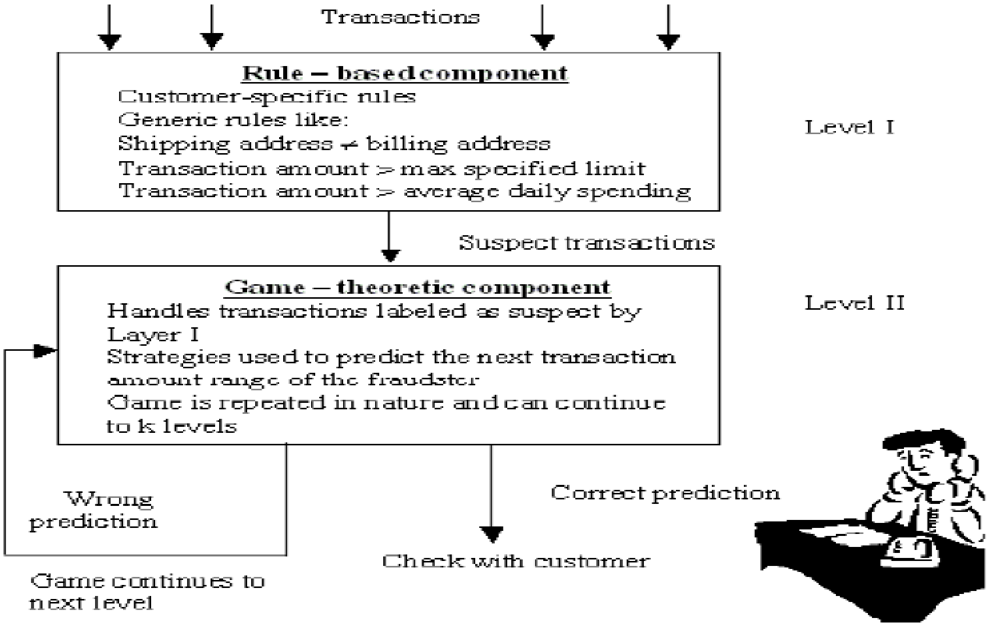

Transactions
Rule – based component
Customer-specific rules
Generic rules like:
Shipping address ≠ billing address
Transaction amount > max specified limit
Transaction amount > average daily spending
Level I
Suspect transactions
Game – theoretic component
Handles transactions labeled as suspect by Layer I
Strategies used to predict the next transaction amount range of the fraudster
Game is repeated in nature and can continue to k levels
Level II
Wrong prediction
Game continues to next level
Correct prediction
Check with customer

Fig 2.5 Architecture of the fraud detection system

The flow of events as would occur in the FDS have been depicted in Figure 2.6.

The transaction for a particular card number is checked at Layer I. If it clears the checks at Level I, it is logged in the master database, failing which it is passed to the Game-theoretic component and the card is marked as suspect. This signifies the beginning of the game between the thief and the FDS.

Layer II predicts the next move of the thief and in the event of the prediction being correct, the card is declared as caught.

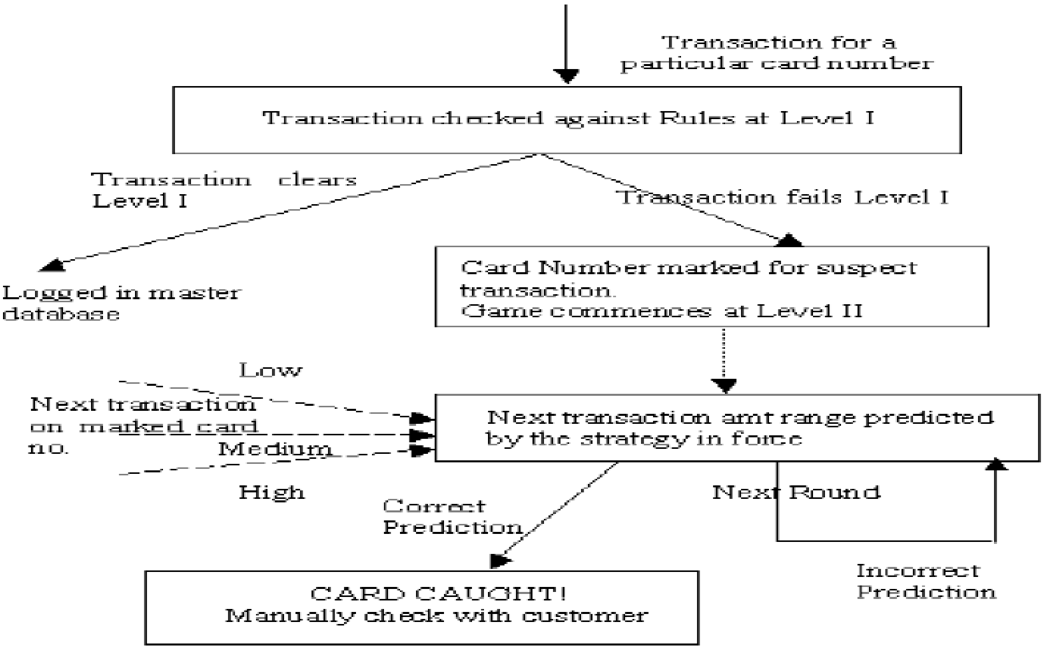

Transaction for a particular card number
Transaction checked against Rules at Level I
Transaction clears Level I
Transaction fails Level I
Logged in master database
Card Number marked for suspect transaction.
Game commences at Level II
Low
Next transaction on marked card no.
Medium
High
Next transaction amt range predicted by the strategy in force
Next Round
Correct Prediction
Incorrect Prediction
CARD CAUGHT!
Manually check with customer

Fig 2.6 Flow of events

2.3.2. Advantages

- Though learning is slower with complex strategies, it does take place in a multi-stage game.

2.3.3. Disadvantages

- The fraudster may eventually learn the methodology being employed.

- Approach is not strategy-specific and other heuristic game-theoretic strategies can be included to further improvise the system.

2.4. Minority Report in Fraud Detection: Classification of Skewed Data [19]

2.4.1. Existing Fraud Detection Techniques

2.4.1.1. Insurance Fraud

Dynamic real-time Bayesian Belief Networks (BBNs), named Mass Detection Tool (MDT) were used for the early detection of potentially fraudulent claims which is then used by a rule generator named Suspicion Building Tool (SBT). The weights of the BBN are refined by the rule generator's outcomes and claim handlers have to keep pace with evolving frauds. This approach evolved from ethnology studies of large insurance companies and loss adjustors who argued against the manual detection of fraud by claim handlers.

The hot spot methodology [28] applies a three step process: the k-means algorithm for cluster detection, the C4.5 algorithm for decision tree rule induction, and domain knowledge, statistical summaries and visualization tools for rule evaluation. It has been applied to detect health care fraud by doctors and the public for the Australian Health Insurance Commission. [29] has expanded the hot spot architecture to use genetic algorithms to generate rules and to allow the domain user, such as a fraud specialist, to explore the rules and to allow them to evolve according to how interesting the discovery is. [30] presented a similar methodology utilizing the Self Organizing Map (SOM) for cluster detection before BP neural networks in automobile injury claims fraud.

Supervised learning is used with BP neural networks, followed by unsupervised learning using SOM to analyze the classification results. Results from clustering show that, out of the four output classification categories used to rate medical practice profiles, only two of the well-defined categories are important. Like the hotspot methodology, this innovative approach was applied on instances of the Australian Health Insurance Commission health practitioners' profiles.

2.4.1.2. Credit card fraud

Network (ANN) comparison study [31] uses the STAGE algorithm for BBNs and BP algorithm for ANNs in fraud detection. Comparative results show that BBNs were more accurate and much faster to train, but BBNs are slower when applied to new instances. Real world credit card data was used but the number of instances is unknown.

The distributed data mining model [32] is a scalable, supervised black box approach that uses a realistic cost model to evaluate C4.5, CART, Ripper and NB classification models. The results demonstrated that partitioning a large data set into smaller subsets to generate classifiers using different algorithms, experimenting with fraud: legal distributions within training data and using stacking to combine multiple models significantly improves cost savings.

This method was applied to one million credit card transactions from two major US banks, Chase Bank and First Union Bank. FairIsaac, formerly known as HNC, produces software for detecting credit card fraud. It favors a three-layer BP neural network for processing transactional, cardholder, and merchant data to detect fraudulent activity.

2.4.1.3. Telecommunications Fraud

The Advanced Security for Personal Communications Technologies (ASPECT) research group [33] focuses on neural networks, particularly unsupervised ones, to train legal current user profiles that store recent user information and user profile histories that store long term information to define normal patterns of use. Once trained, fraud is highly probable when there is a difference between a mobile phone user's current profile and the profile history.

The adaptive fraud detection framework presents rule-learning fraud detectors based on account-specific thresholds that are automatically generated for profiling the fraud in an individual account. The system, based on the framework, has been applied by combining the most relevant rules, to uncover fraudulent usage that is added to the legitimate use of a mobile phone account.

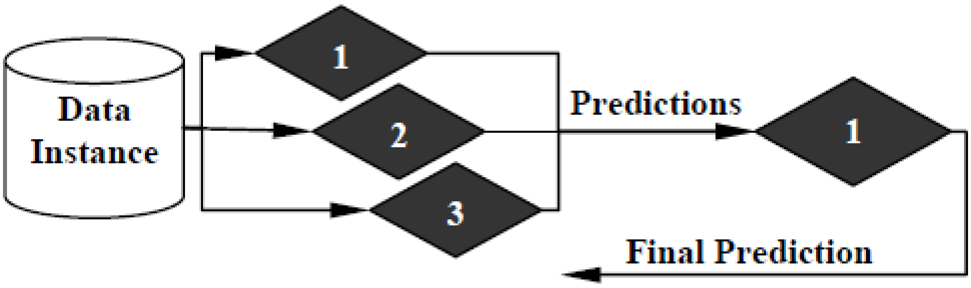


Fig 2.7 Predictions on a single data instance using precogs

2.4.2. The new Fraud Detection Method

The idea is to simulate the book's [34] Precrime method of precogs and integration mechanisms with existing data mining methods and techniques. An overview of how the new method can be used to predict fraud for each instance is provided.

Precogs, or precognitive elements, are entities that have the knowledge to predict that something will happen.

Figure 2.7 shows that as each precog output its many predictions for each instance, all the predictions are fed back into one of the precogs, to derive a final prediction for each instance.

2.4.3. Fraud Detection Algorithms

2.4.3.1. Classifiers

- Although the naive Bayesian (NB) algorithm is simple, it is very effective in many real world data sets because it can give better predictive accuracy than well-known methods like C4.5 and BP and is extremely efficient in that it learns in a linear fashion using ensemble mechanisms, such as bagging and boosting, to combine classifier predictions. However, when attributes are redundant and not normally distributed, the predictive accuracy is reduced.

- C4.5 can help not only to make accurate predictions from the data but also to explain the patterns in it. It deals with the problems of the numeric attributes, missing values, pruning, estimating error rates, complexity of decision tree induction, and generating rules from trees. However, scalability and efficiency problems, such as the substantial decrease in performance and poor use of available system resources, can occur when C4.5 is applied to large data sets.

- Back propagation (BP) neural networks can process a very large number of instances; have a high tolerance to noisy data; and has the ability to classify patterns which they have not been trained. They are an appropriate choice if the results of the model are more important than understanding how it works. However, the BP algorithm requires long training times and extensive testing and retraining of parameters, such as the number of hidden neurons, learning rate and momentum, to determine the best performance.

Algorithm	Effectiveness	Scalability	Speed
NB	Good	Excellent	Excellent
C4.5	Excellent	Poor	Good
BP	Good	Excellent	Poor

Table 2.1 Qualitative comparison of classifiers

2.4.3.2. Combining Outputs

- Bagging combines the classifiers trained by the same algorithm using unweighted majority voting on each example or instance. Voting denotes the contribution of a single vote, or its own prediction, from a classifier. The final prediction is then decided by the majority of the votes. Generally, bagging performs significantly better than the single model for C4.5 and BP algorithms. It is never substantially worse because it neutralizes the instability of the classifiers by increasing the success rate.

- Stacking combines multiple classifiers generated by different algorithms with a meta-classifier. To classify an instance, the base classifiers from the three algorithms present their predictions to the meta-classifier which then makes the final prediction.

- Stacking-bagging is a hybrid technique proposed by this paper. The recommendation here is to train the simplest learning algorithm first, followed by the complex ones. In this way, NB base classifiers are computed, followed by the C4.5 and then the BP base classifiers. The NB predictions can be quickly obtained and analyzed while the other predictions, which take longer training and scoring times, are being processed.

As most of the classification work has been done by the base classifiers, the NB algorithm, which is simple and fast, is used as the meta-classifier. In order to select the most reliable base classifiers, stacking-bagging uses stacking to learn the relationship between classifier predictions and the correct class. For a data instance, these chosen base classifiers' predictions then contribute their individual votes and the class with the most votes is the final prediction.

2.5. A Neural Network Based Database Mining System for Credit Card Fraud Detection [35]

2.5.1. Architecture

Today's data mining tools have typically evolved out of the pattern recognition and artificial intelligence research. These tools have a heavy algorithmic component and are often rather "bare" with respect to user friendliness and generality. They mostly work on flat files, which imposes a significant constraint for their deployment in a corporate environment. In case of modern corporate databases, copying huge data sets from databases into flat files is not tolerable. The possibility to directly access different database types becomes a critical requirement for modern database mining systems. Furthermore, a sophisticated yet straightforward graphical user interface (GUI) is a must. All these requirements are provided by CARDWATCH.

This system consists of five main modules:

- **Global Constants Module (GCM)**: The purpose of this module is to bundle all the global variables declared in the system (except the external dynamic link library (DLL) part). For example, record sets used to train and test the detection algorithms are globally declared in the GCM and accessed by other modules. The GCM is implemented in Visual Basic.

- **Core/Graphical User Interface Module (GUIM):** This module not only allows the user to comfortably control the entire system, but also serves as the "glue" for all other modules. It serves as a container for all GUI-related routines, including the callback code or auxiliary functions for widget control. Moreover, this modules handles the creation of neural network description files, which are then accessed by the LAIM module and forwarded to the LAL module. The GUIM communicates with the LAIM, DBIM and GCM module. It is implemented in Visual Basic.

- **Database Interface Module (DBIM):** This module handles the communication between the database and the remaining modules. It contains the code for such operations as initialization, opening and modification of databases, assignment of database fields to GUI data control widgets, querying of the databases via SQL, or assignment of selected record sets to the global variables. Currently, the database systems Microsoft Access, dBase, FoxPro, Paradox and ODBC compatible systems are supported; the test version for the credit card application uses MS Access. The DBIM cooperates with the GUIM, GCM and LAIM module. It is implemented in Visual Basic with inline SQL statements.

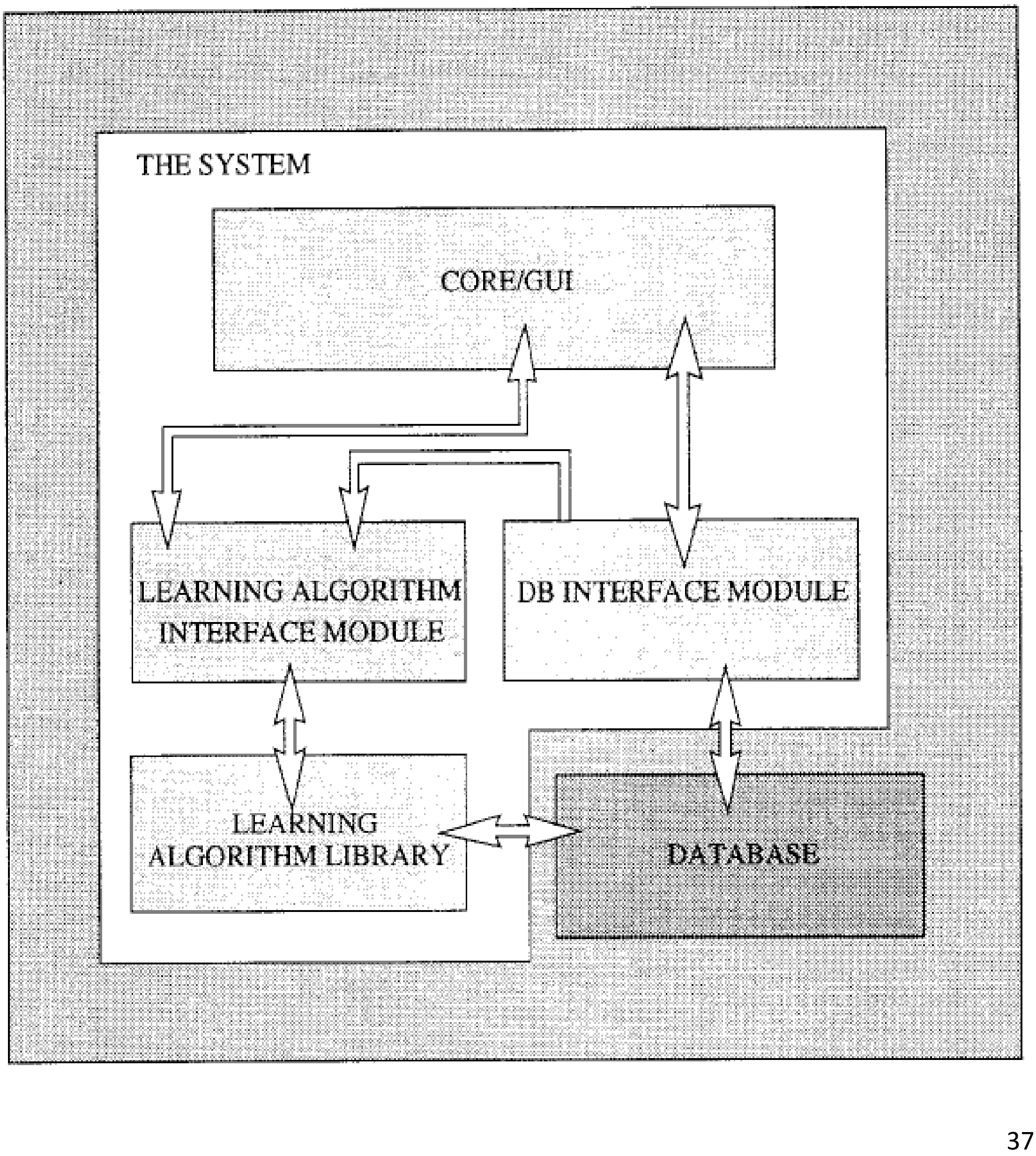

THE SYSTEM
CORE/GUI
LEARNING ALGORITHM INTERFACE MODULE
DB INTERFACE MODULE
LEARNING ALGORITHM LIBRARY
DATABASE

Fig 2.8 Architecture of CARDWATCH

- **Learning Algorithms Library (LAL):** This module provides the neural network learning algorithms. In the current version, it is limited to only a few neural network architectures with three different learning rules, but it is easily extensible to any other adaptive techniques which are used to detect anomalies in a customer's credit card usage dynamics. The module has own database access facilities making it autonomous, i.e. independent of the core part of the system while retrieving transaction data or marking fraudulent records.

  This keeps the interfaces to the core highly efficient. No transaction data interchange takes place at the boundaries of the LAIM and LAL modules. The only information required by the LAL is the name of the corresponding database, its type, the name of the network description file, the name of the network optimization parameters description file, and a few arguments of simple data types.

  There is a number of routines made available to the core part of CARDWATCH including DLL initialization functions, database initialization functions, learning and detection functions, or a routine for saving network parameters. The LAIM module communicates with the LAL module and the database. It is implemented using the SINElib neural library in form of a Visual C++ dynamic link library (DLL).

- **Dynamic link library (DLL):** This module provides an interface between the core and the neural network library. This is a rather small module containing a simple interface to external LAL functions. There are only two functions inside: train and test with method dependent calls to LAL test/train functions. The method dependency is aimed to provide an option for future enhancement of CARDWATCH by further adaptive methods.

2.6. Credit Card Fraud Detection Using Genetic Algorithm [36][37]

2.6.1. Genetic Algorithm

Genetic Algorithm (GA) is an optimization technique that attempts to replicate natural evolution processes in which the individuals with the considered best characteristics to adapt to the environment are more likely to reproduce and survive. These advantageous individuals mate between them, producing descendants similarly characterized, so favorable characteristics are preserved and unfavorable ones destroyed, leading to a progressive evolution of the species.

In other words, the basic idea of genetic algorithms is that given a problem, the genetic pool of a specific population potentially contains the solution, or a better solution. Based on genetic and evolutionary principles, the genetic algorithm repeatedly modifies a population of artificial structures through the application of initialization, selection, crossover, and mutation operators in order to obtain an evolved solution.

Artificial genetic algorithm aims to improve the solution to a problem by keeping the best combination of input variables. It starts with the definition of the problem to optimize, generating an objective function to evaluate the possible candidate solutions (chromosomes), i.e., the objective function is the way of determining which individual produces the best outcome.

2.6.2. Algorithm

Step1: Input group of data credit card transactions, every transaction record with n attributes, and standardize the data, get the sample finally, which includes the confidential information about the card holder, store in the data set.

Step2: Compute the critical values, Calculate the CC usage frequency count, CC usage location, CC overdraft, current bank balance, average daily spending

Step3: Generate critical values found after limited number of generations. Critical Fraud Detected, Monitor able Fraud Detected, Ordinary Fraud Detected etc. using Genetic algorithm

Step4: Generate fraud transactions using this algorithm. This is to analyze the feasibility of credit card fraud detection based on technique, applies detection mining based on critical values into credit card fraud detection and proposes this detection procedures and its process.

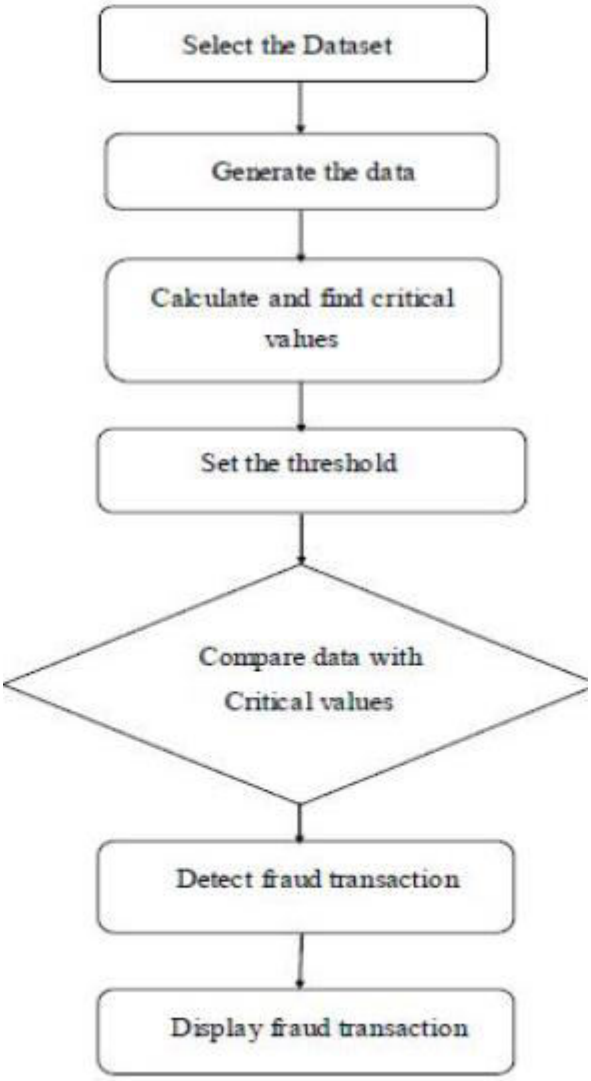

Select the Dataset
Generate the data
Calculate and find critical values
Set the threshold
Compare data with Critical values
Detect fraud transaction
Display fraud transaction

Fig 2.9 Flow of Genetic Algorithm Process

2.7. Comparison of various Techniques

Method	Technique	Processing Speed	Cost	Accuracy	Research issues addressed	Research Challenges
Genetic Programming, algorithms	The Evolutionary-Fuzzy System- A GP Approach	Low	Implementation is highly expensive	Very High	Easily detect stolen credit card Frauds. Detect suspicious, non-suspicious data	Not applicable in E-Commerce, Difficult to implement
Neural Networks	ANN & BNN	Low	Expensive	Medium	Cellular phone fraud, Calling card fraud, Computer Network Intrusion Applicable in E-Commerce	Needs training to operate and requires high processing time for large neural networks and BNN
Behavioural Analysis	Hidden Markov Model	High	Quite expensive	Medium	Applicable in online detection of credit card fraud.	High false alarm,
Clustering	Peer-Group Analysis, Break-point Analysis	Low	Expensive	High	The original user is not checked as it maintains a log	False Positive is high
Decision Tree	Similarity Tree	Low	Low Expensive	High	Identify suspicious data	High false alarm

Table 2.2 Comparison of various algorithms

Chapter 3

Proposed Work & Implementation

Proposed Work & Implementation

A credit cardholder makes different kinds of purchases of different amounts over a period of time. One possibility is to consider the sequence of transaction amounts and look for deviations in them. However, the sequence of types of purchase is more stable compared to the sequence of transaction amounts. The reason is that, a cardholder makes purchases depending on his need for procuring different types of items over a period of time. This, in turn, generates a sequence of transaction amounts.

Each individual transaction amount usually depends on the corresponding type of purchase. Hence, we consider the transition in the type of purchase as state transition in our model. The type of each purchase is linked to the line of business of the corresponding merchant. This information about the merchant's line of business is not known to the issuing bank running the FDS. Thus, the type of purchase of the cardholder is hidden from the FDS.

The set of all possible types of purchase and, equivalently, the set of all possible lines of business of merchants forms the set of hidden states of the HMM. It should be noted at this stage that the line of business of the merchant is known to the acquiring bank, since this information is furnished at the time of registration of a merchant. Also, some merchants may be dealing in various types of commodities (For example, Wal-Mart, K-Mart, or Target sells tens of thousands of different items).

Such types of line of business are considered as Miscellaneous, and we do not attempt to determine the actual types of items purchased in these transactions. Any assumption about availability of this information with the issuing bank and, hence, with the FDS, is not practical and, therefore, would not have been valid.

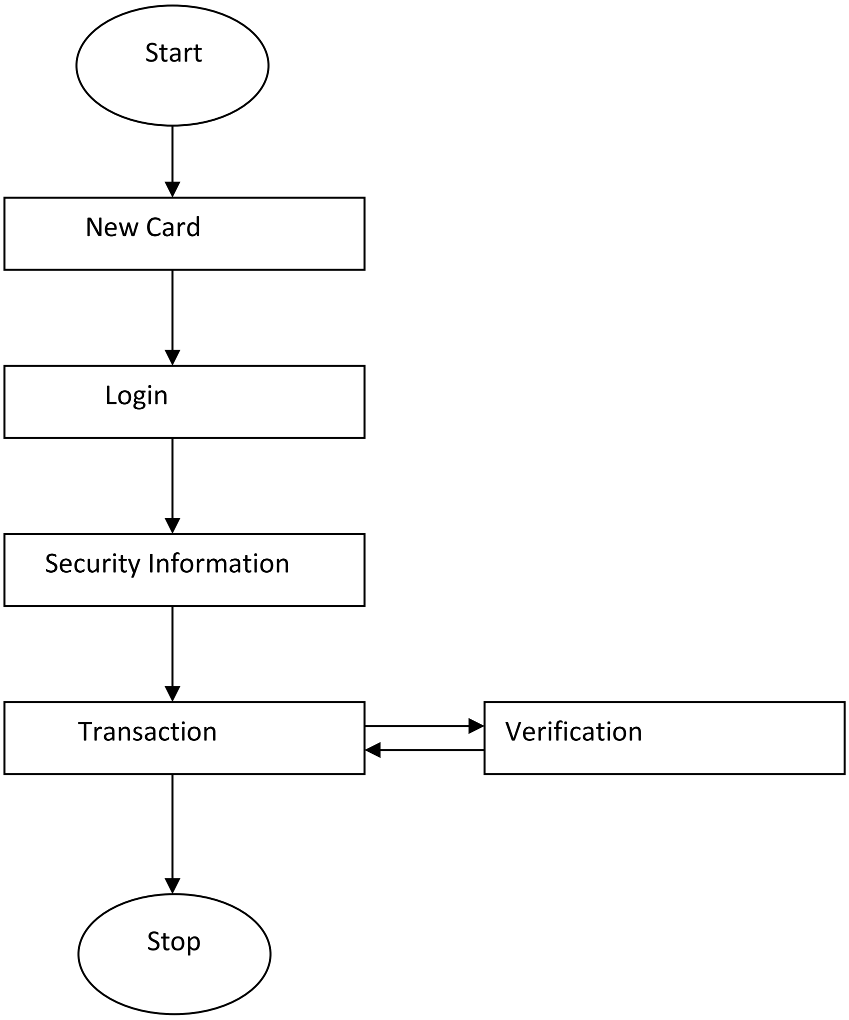


Fig 3.1 Proposed System

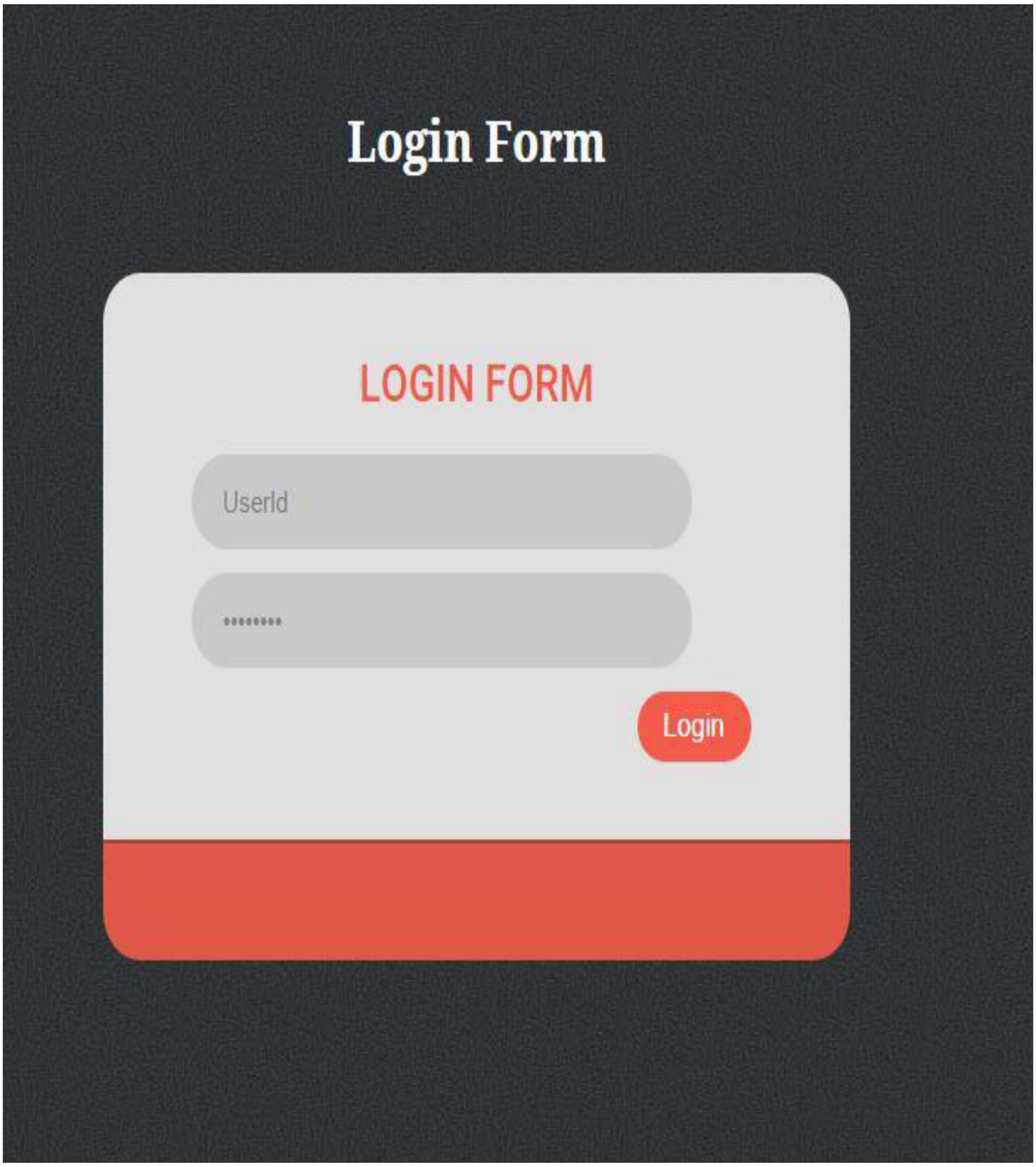
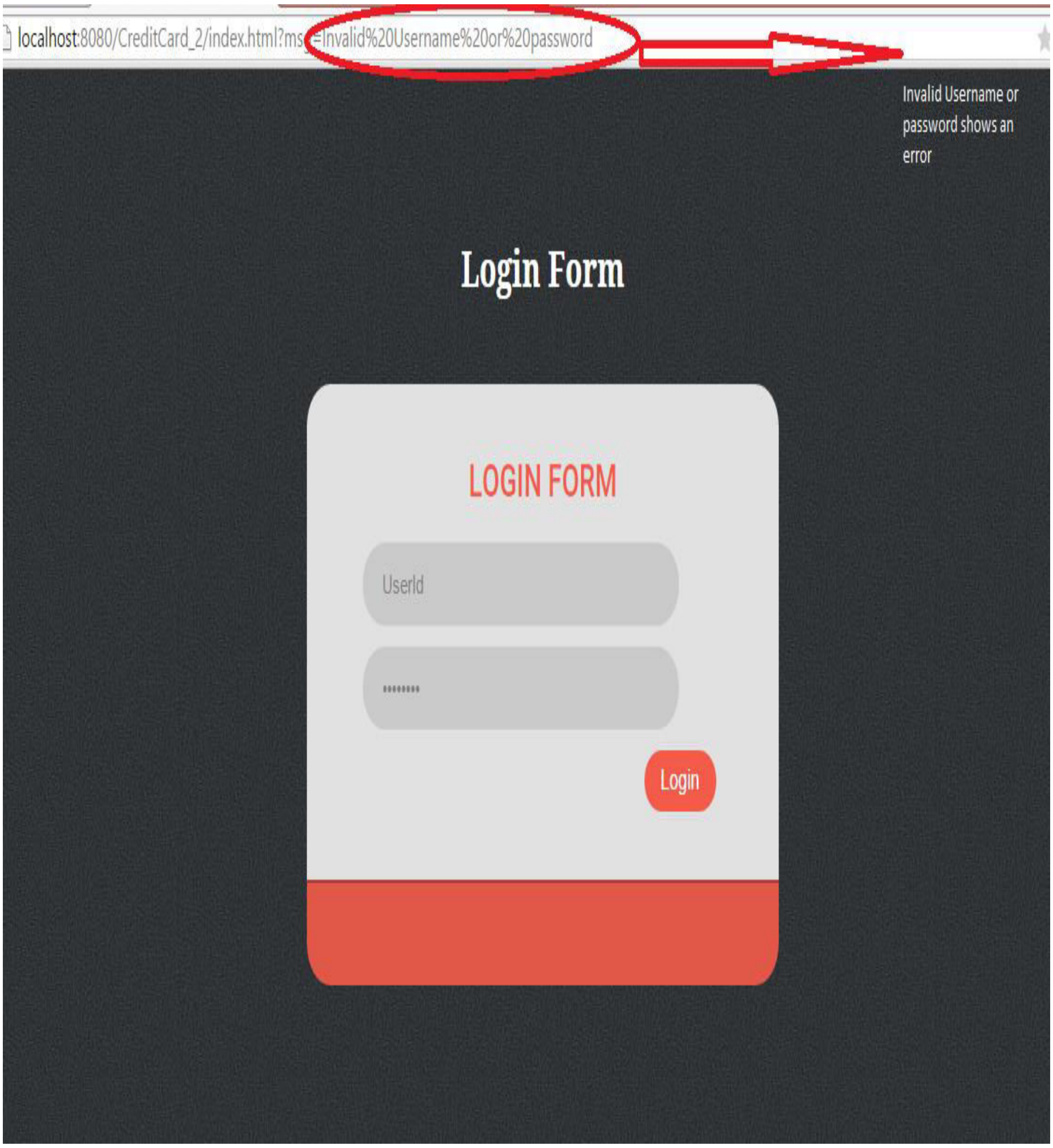


Fig 3.2 Login Form

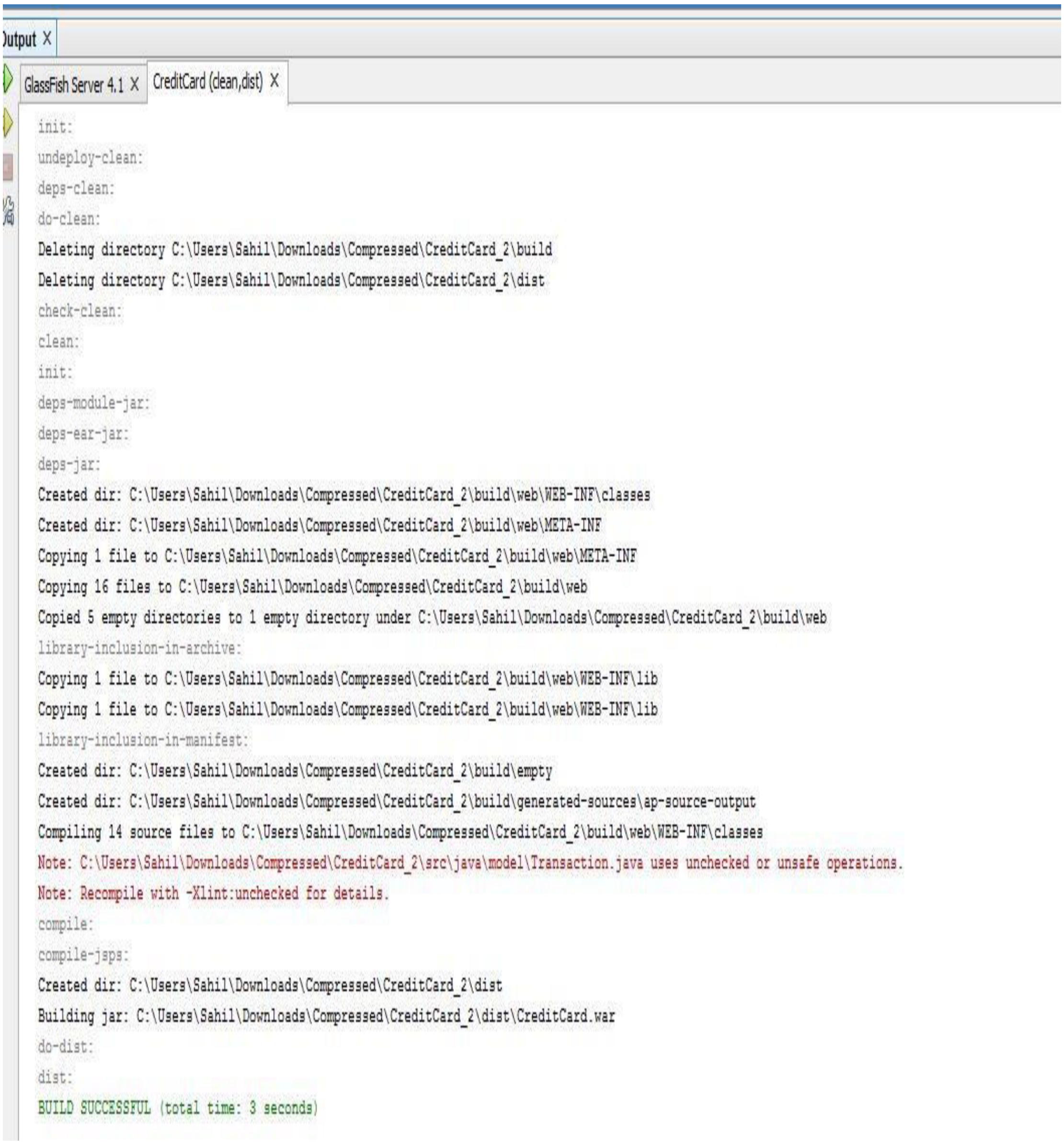


localhost:8080/CreditCard_2/index.html?ms=Invalid%20Username%20or%20password

Invalid Username or password shows an error

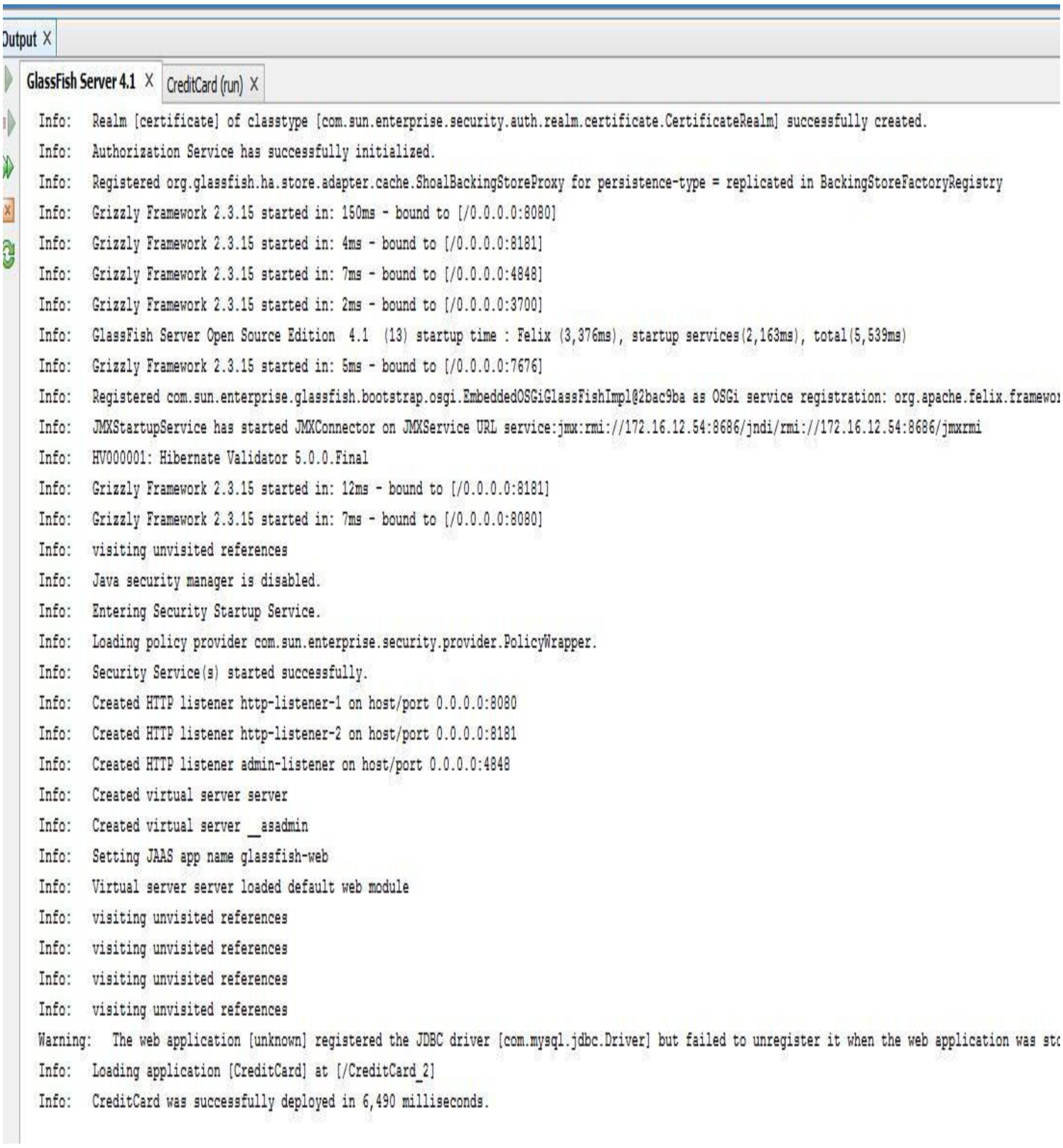
Login Form

Fig 3.3 Login form showing error for invalid user



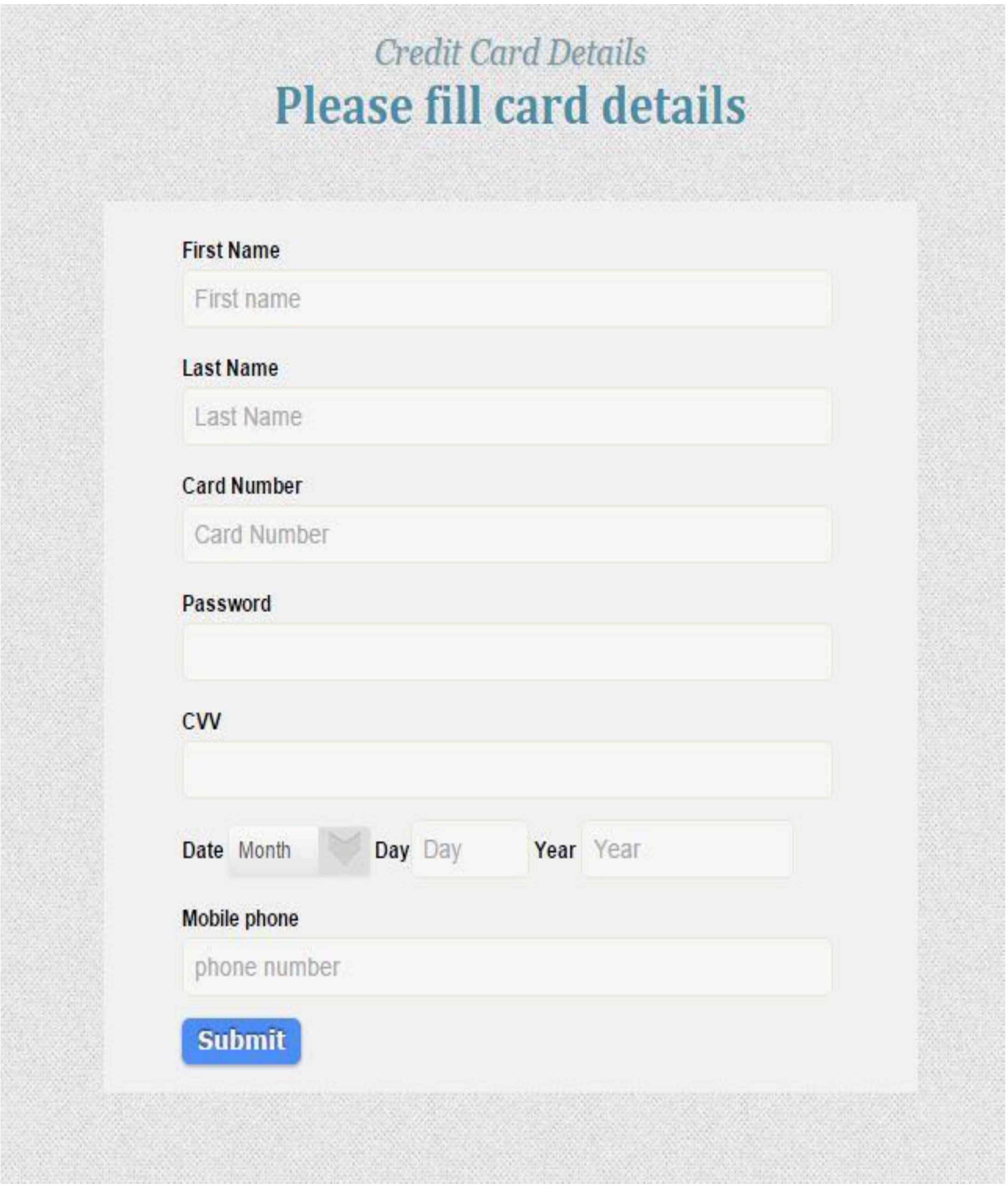
```
init:
undeploy-clean:
deps-clean:
do-clean:
Deleting directory C:\Users\Sahil\Downloads\Compressed\CreditCard_2\build
Deleting directory C:\Users\Sahil\Downloads\Compressed\CreditCard_2\dist
check-clean:
clean:
init:
deps-module-jar:
deps-ear-jar:
deps-jar:
Created dir: C:\Users\Sahil\Downloads\Compressed\CreditCard_2\build\web\WEB-INF\classes
Created dir: C:\Users\Sahil\Downloads\Compressed\CreditCard_2\build\web\META-INF
Copying 1 file to C:\Users\Sahil\Downloads\Compressed\CreditCard_2\build\web\META-INF
Copying 16 files to C:\Users\Sahil\Downloads\Compressed\CreditCard_2\build\web
Copied 5 empty directories to 1 empty directory under C:\Users\Sahil\Downloads\Compressed\CreditCard_2\build\web
library-inclusion-in-archive:
Copying 1 file to C:\Users\Sahil\Downloads\Compressed\CreditCard_2\build\web\WEB-INF\lib
Copying 1 file to C:\Users\Sahil\Downloads\Compressed\CreditCard_2\build\web\WEB-INF\lib
library-inclusion-in-manifest:
Created dir: C:\Users\Sahil\Downloads\Compressed\CreditCard_2\build\empty
Created dir: C:\Users\Sahil\Downloads\Compressed\CreditCard_2\build\generated-sources\ap-source-output
Compiling 14 source files to C:\Users\Sahil\Downloads\Compressed\CreditCard_2\build\web\WEB-INF\classes
Note: C:\Users\Sahil\Downloads\Compressed\CreditCard_2\src\java\model\Transaction.java uses unchecked or unsafe operations.
Note: Recompile with -Xlint:unchecked for details.
compile:
compile-jsps:
Created dir: C:\Users\Sahil\Downloads\Compressed\CreditCard_2\dist
Building jar: C:\Users\Sahil\Downloads\Compressed\CreditCard_2\dist\CreditCard.war
do-dist:
dist:
BUILD SUCCESSFUL (total time: 3 seconds)
```

Fig 3.4 Building the system



```
Info:   Realm [certificate] of classtype [com.sun.enterprise.security.auth.realm.certificate.CertificateRealm] successfully created.
Info:   Authorization Service has successfully initialized.
Info:   Registered org.glassfish.ha.store.adapter.cache.ShoalBackingStoreProxy for persistence-type = replicated in BackingStoreFactoryRegistry
Info:   Grizzly Framework 2.3.15 started in: 150ms - bound to [/0.0.0.0:8080]
Info:   Grizzly Framework 2.3.15 started in: 4ms - bound to [/0.0.0.0:8181]
Info:   Grizzly Framework 2.3.15 started in: 7ms - bound to [/0.0.0.0:4848]
Info:   Grizzly Framework 2.3.15 started in: 2ms - bound to [/0.0.0.0:3700]
Info:   GlassFish Server Open Source Edition  4.1  (13) startup time : Felix (3,376ms), startup services(2,163ms), total(5,539ms)
Info:   Grizzly Framework 2.3.15 started in: 5ms - bound to [/0.0.0.0:7676]
Info:   Registered com.sun.enterprise.glassfish.bootstrap.osgi.EmbeddedOSGiGlassFishImpl@2bac9ba as OSGi service registration: org.apache.felix.framewo
Info:   JMXStartupService has started JMXConnector on JMXService URL service:jmx:rmi://172.16.12.54:8686/jndi/rmi://172.16.12.54:8686/jmxrmi
Info:   HV000001: Hibernate Validator 5.0.0.Final
Info:   Grizzly Framework 2.3.15 started in: 12ms - bound to [/0.0.0.0:8181]
Info:   Grizzly Framework 2.3.15 started in: 7ms - bound to [/0.0.0.0:8080]
Info:   visiting unvisited references
Info:   Java security manager is disabled.
Info:   Entering Security Startup Service.
Info:   Loading policy provider com.sun.enterprise.security.provider.PolicyWrapper.
Info:   Security Service(s) started successfully.
Info:   Created HTTP listener http-listener-1 on host/port 0.0.0.0:8080
Info:   Created HTTP listener http-listener-2 on host/port 0.0.0.0:8181
Info:   Created HTTP listener admin-listener on host/port 0.0.0.0:4848
Info:   Created virtual server server
Info:   Created virtual server __asadmin
Info:   Setting JAAS app name glassfish-web
Info:   Virtual server server loaded default web module
Info:   visiting unvisited references
Info:   visiting unvisited references
Info:   visiting unvisited references
Info:   visiting unvisited references
Warning:   The web application [unknown] registered the JDBC driver [com.mysql.jdbc.Driver] but failed to unregister it when the web application was st
Info:   Loading application [CreditCard] at [/CreditCard_2]
Info:   CreditCard was successfully deployed in 6,490 milliseconds.
```

Fig 3.5 Running glassfish server



Credit Card Details

Please fill card details

First Name

First name

Last Name

Last Name

Card Number

Card Number

Password

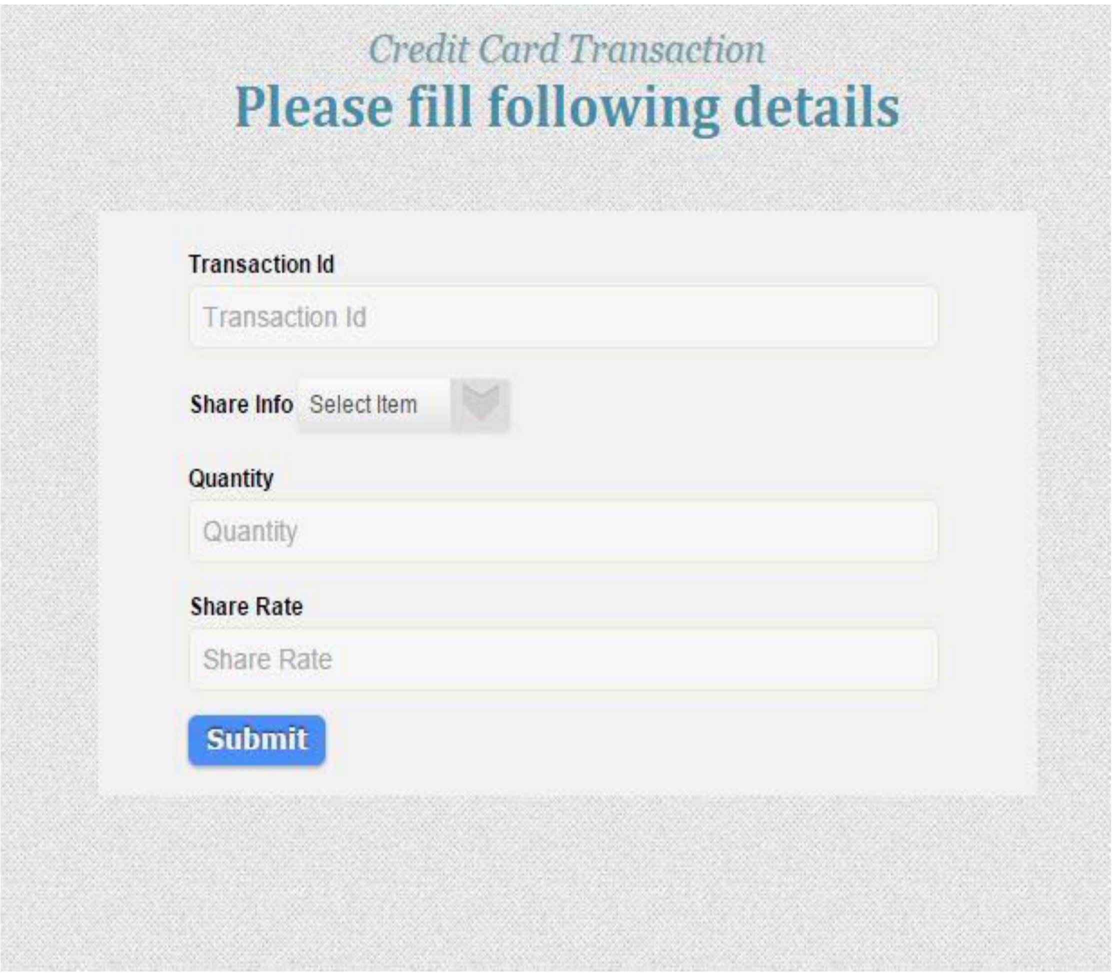
CVV

Date Month Day Day Year Year

Mobile phone

phone number

Fig 3.6 Form for filling the card information



Credit Card Transaction

Please fill following details

Transaction Id

Transaction Id

Share Info Select Item

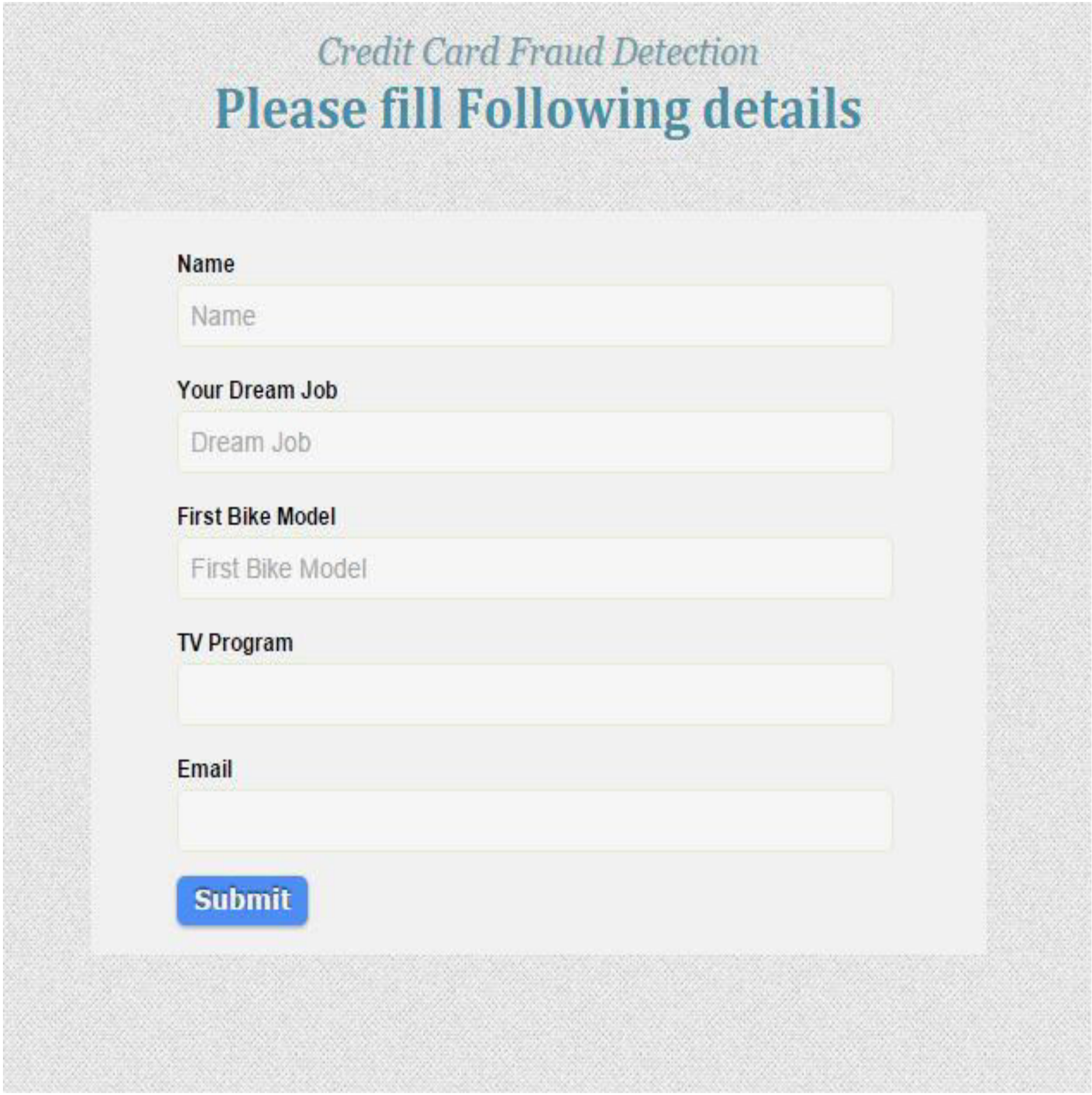
Quantity

Quantity

Share Rate

Share Rate

Fig 3.7 Form for filling the transaction information



Credit Card Fraud Detection

Please fill Following details

Name

Your Dream Job

First Bike Model

TV Program

Email

Fig 3.8 Form for filling security questions

Chapter 4

Results & Conclusion

Chapter 4

Results & Conclusion

4.1. Comparison between Naïve Bayesian and Back Propagation

	Execution Time	**Accuracy**
Naïve Bayesian	5 s	54%
Back Propagation	9 s	61%

Table 4.1 Comparison between NB and BP

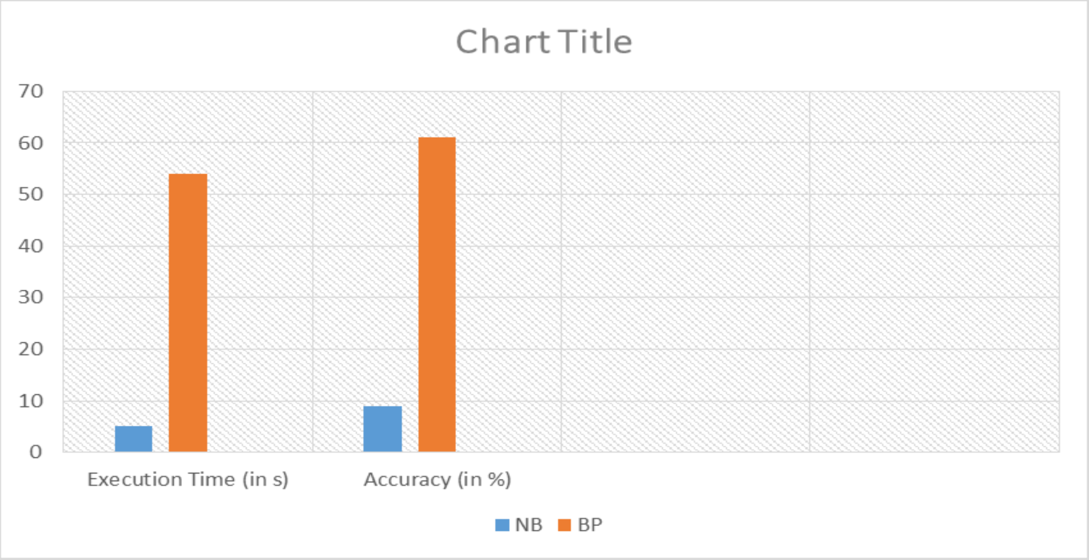


Fig 4.1 Comparison between NB and BP

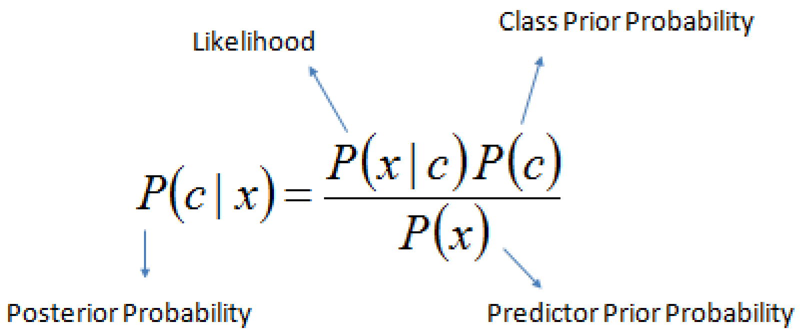
Naïve Bayesian: Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a tuple belongs to a particular class. It is based on Bayes' theorem. [38]

Let 'X' be a data tuple. In Bayesian terms, X is evidence. As usual, it is described by measurements made on the set of 'n' attributes. Let 'H' be some hypothesis such that the data tuple X belongs to specifies class 'C'. For classification problems, we want to determine P(H/X). The problem that the hypothesis H holds given the evidence or observer data tuple. In other words, we are looking for the problem that the tuple X belongs to class C given that we know the attribute description of X.

P(H/X) is the posterior probability of H conditioned on X. In contrast, P(H) is the prior probability of H. The posterior probability is based on more information than the prior probability which is independent of X. Similarly, P(X/H) is the posterior probability of X conditioned on H. P(X) is the prior probability of X.

$$P(H/X) = \frac{P(X/H)P(H)}{P(X)}$$

Assumption is that attributes are independent (it makes computation simple). That is why it is called Naïve Bayesian Classification.


Likelihood
Class Prior Probability
$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Back propagation: BP learns by iteratively processing a dataset of training tuples by comparing the network's prediction for each tuple with the actual known target value. The target value may be the known class label of the training tuple. For each training tuple, the weights are modified so as to minimize the mean squared error between the network's prediction and the actual target value. [39]

The modifications are made in the backwards direction i.e. from the output layer through each hidden layer down to the first layer. Hence, the name Back Propagation. Although it is not guaranteed, in general the weights will eventually converge and the learning process stops.

4.2. Conclusion

Credit cards have become a great source of money for fraudsters. This is due to the increase in the use of credit cards. Everyone prefers using credit cards because it makes our life easy. To decrease frauds in credit cards we have proposed a credit card fraud detection system using HMM and Stochastic tools & Technology.

We compared both the techniques and found that stochastic proves better in terms of accuracy. We also compared the accuracy and execution time of NB and BP algorithm.

References

[1] Azeem Ush Shan Khan, Nadeem Akhtar and Mohammad Naved Qureshi, ‘Real-Time Credit-Card Fraud Detection using Artificial Neural Network Tuned by Simulated Annealing Algorithm’, Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC, Association of Computer Electronics and Electrical Engineers, 2014.

[2] Yusuf Sahin, Serol Bulkan, Ekrem Duman, ‘A cost-sensitive decision tree approach for fraud detection’, Expert Systems with Applications, Elsevier, 2013.

[3] W. Roberds, ‘The impact of fraud on new methods of retail payment, Federal Reserve Bank of Atlanta Economic Review’, First Quarter (1998) 42–52.

[4] Statistics for General and Online Card Fraud, 20 June, 2007. <http://epaynews.com/statistics/fraud.html>.

[5] Online fraud is 12 times higher than offline fraud, 20 June, 2007. <http://sellitontheweb.com/ezine/news0434.shtml>.

[6] Suvasini Panigrahi, Amlan Kundu, Shamik Sural, A.K. Majumdar, ‘Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning’, Information Fusion 10, Elsevier, page 354–363, 2009.

[7] Kavita Rawat ,Jyoti Hazrati, ‘Credit card Fraud detection using Hidden Markov Model’, International Journal of Latest Research in Science and Technology ISSN (Online):2278-5299 Vol.1,Issue 4, Page No.420-422 ,November-December 2012.

[8] “Consumer Sentinel Network Data Book: January – December 2008”, Federal Trade Commission, 26 February 2009, Retrieved 21 February 2010.

[9] “Court filings double estimate of TJX breach”, 2007.

[10] Adsit, Dennis, “Error-proofing strategies for managing call center fraud”, isixsigma.com, 21 February 2011.

[11] http://en.wikipedia.org/wiki/Credit_card_fraud

[12] Inside Job/Restaurant card skimming. Journal Register.

[13] Little, Allan, “Overseas credit card scam exposed”, bbc.co.uk.com, 19 March 2009.

[14] NACS Magazine – Skimming, nacsonline.com

[15]http://www.ukessays.com/essays/information-technology/credit-card-fraud-types-and-detection-methods-information-technology-essay.php

[16] Nitin Mishra, Ranjit Kumar, Shishir Kumar Shandilya, “Credit Card Fraud Transaction Detection by using Hidden Markov Model”, International Journal of Scientific Engineering and Technology, Vol. 1, Issue 2, page: 139-142, 2012.

[17] C. Phua, V. Lee, K. Smith, and R. Gayler, “A Comprehensive Survey of Data Mining-Based Fraud Detection Research,” http://w.bsys.monash.edu.au/people/cphua/, Mar. 2007.

[18] S. Stolfo and A.L. Prodromidis, “Agent-Based Distributed Learn-ing Applied to Fraud Detection,” Technical Report CUCS-014-99, Columbia Univ., 1999.

[19] C. Phua, D. Alahakoon, and V. Lee, “Minority Report in Fraud Detection: Classification of Skewed Data,” ACM SIGKDD Explora-tions Newsletter, vol. 6, no. 1, pp. 50-59, 2004.

[20] V. Vatsa, S. Sural, and A.K. Majumdar, “A Game-theoretic Approach to Credit Card Fraud Detection,” Proc. First Int’l Conf. Information Systems Security, pp. 263-276, 2005.

[21] S. Axelsson, “The Base-Rate Fallacy and the Difficulty of Intrusion Detection,” ACM Trans. Information and System Security, vol. 3, no. 3, pp. 186-205, 2000.

[22] L.R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proc. IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[23] C. Chiu and C. Tsai, “A Web Services-Based Collaborative Scheme for Credit Card Fraud Detection,”Proc. IEEE Int’l Conf. e-Technology, e-Commerce and e-Service, pp. 177-181, 2004.

[24] XML Schema, http://www.w3.org/xml/schema

[25] Predictive Model Markup Language, http://www.dmg.org/pmml-v2-0.htm

[26] Document Type Definition, http://www.w3schools.com/dtd

[27] R. Agrawal, T. Imielinski, and A.N. Swami, “Mining Association Rules Between Sets of Items in Large Databases”, Proc. of the 1993 ACM SIGMOD Int. Conf. on Management of Data, pp. 207-216, 1993.

[28] Williams G and Huang Z. “Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large Real World Databases”, in Proceedings of the 10th Australian Joint Conference on Artificial Intelligence, Perth, Australia, 1997.

[29] Williams G. “Evolutionary Hot Spots Data Mining: An Architecture for Exploring for Interesting Discoveries”, in Proceedings of the 3rd Pacific-Asia Conference in Knowledge Discovery and Data Mining, Beijing, China, 1999.

[30] Brockett P, Xia X and Derrig R. “Using Kohonen's Self Organising Feature Map to Uncover Automobile Bodily Injury Claims Fraud”, Journal of Risk and Insurance, USA, 1998.

[31] Maes S, Tuyls K, Vanschoenwinkel B and Manderick B. “Credit Card Fraud Detection Using Bayesian and Neural Networks”, in Proceedings of the 1st International NAISO Congress on Neuro Fuzzy Technologies, Havana, Cuba, 2002.

[32] Provost F. “Machine Learning from Imbalanced Data Sets 101”, Invited paper, in Workshop on Learning from Imbalanced Data Sets, AAAI, Texas, USA, 2000.

[33] Weatherford M. “Mining for Fraud”, IEEE Intelligent Systems, July/August Issue, pp4-6, 2002.

[34] Dick P K. Minority Report, Orion Publishing Group, London, Great Britain, 1956.

[35] E. Aleskerov, B. Freisleben, and B. Rao, “CARDWATCH: A Neural Network Based Database Mining System for Credit Card Fraud Detection,” Proc. IEEE/IAFE: Computational Intelligence for Financial Eng., pp. 220-226, 1997.

[36] Rinky D. Patel, Dheeraj Kumar Singh, ‘Credit Card Fraud Detection & Prevention of Fraud Using Genetic Algorithm’, International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-2, Issue-6, January 2013.

[37] Mayuri Agrawal, Sonali Rangdale, ‘Discovering Fraud in Credit Card by Genetic Programming’, International Journal of Innovative Research & Development, page 130-132, ISSN 2278 – 0211, Vol 3 Issue 11, November, 2014.

[38] http://www.saedsayad.com/naive_bayesian.htm

[39] http://en.wikipedia.org/wiki/Backpropagation